Kernel-based manifold visualization of GPCR sequences

Martha Ivón Cárdenas Domínguez

June 22, 2011

To my sons Ivón, Alan and Laia, my rays of light

Acknowledgements

During the master thesis preparation, I have gotten many good advices. This thesis would not have been possible without the accomplishments and support of many people.

Firstly, I would like to thank my advisors Dr. Alfredo Vellido and Dr. Jesús Giraldo, for giving me the opportunity to study this topic, helping me with questions and providing me useful feedback, being of a very great help for me. Both of them provided me valuable advice and input at every stage of this thesis, spending a long time during the preparation of this work.

Secondly, I would like to thank Dr. Iván Olier from whose work stems my own and who spent many hours supervising the experiment results. Similarly, I would like to thank Dr. Xavier Rovira who helped me to understand the biological problem and provided the sequence data used in my experiments. Their expertise was crucial in the design of the experiments.

My parents, Rafael and Marta, and my family are a source of enormous love and support and have maintained an avid, ongoing interest in my work. My husband, Albert, has cheerfully put up with my crazy schedule.

Finally, I am especially grateful to all the professors of my master degree who helped me, shared their knowledge with me and improved the quality of my work.

Abstract

G-Protein Coupled Receptors (GPCRs) are key players in cell-cell communication. They transduce a wide range of extracellular signals such as light, odors, hormones or neurotransmitters into appropriated cellular responses. These receptors regulate many cell functions and are encoded by the largest gene family in mammalian genomes, representing more than 3% of the human genes. GPCRs are the estimated target of approximately half of the medicines currently in clinical use.

Probabilistic modelling and specifically, machine learning probabilistic models have only recently begun to be applied to the analysis of GPCR functioning, although their application is expected to generate new insights in this field. Statistical machine learning techniques are specially suited to deal with some of the common challenges of molecular modelling in proteins, and should be of special interest when the three dimensional structures of the proteins and receptors remain unknown at large.

In this thesis, we describe a statistical machine learning model of the manifold learning family, adapted through kernelization to the analysis of protein sequence data. Experimental results show that it provides a differentiated visualization and grouping of GPCR

subfamilies and that these groupings faithfully reflect the structure of GPCR phylogenetic trees.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

It has been just over 10 years since the publication of the first draft of the human genome decoding [39]. The detailed description of the human genome is a milestone of science in general and of medicine in particular. It has opened the doors to new approaches to investigate pathologies that hold the promise of the advent of truly personalized medicine. Through these doors, though, a new challenge for intelligent data analysis has also entered.

Over the last decade, medicine has become a data-intensive area of research. One in which new data-acquisition technologies and a wider variety of investigative goals coalesce to make it one of the most important challenges for intelligent data analysis [36]. The *-omics* sciences have contributed the most to this data deluge, stemming from microarrays in genomics, from protein chips and tissue arrays in proteomics, etc. As very explicitly reported in [29] [...] *the need to process terabytes of information has become de rigueur for many labs engaged in genomic research.*

Arguably, drug research has contributed more to the progress of medicine during the past century than any other scientific factor [12]. One of the main areas of drug research deals with the analysis of proteins. The function of the

proteins depends directly on their 3D structure, which is embodied in their amino acid sequence. Such 3D structure is difficult to unravel, but protein sequences are easy to acquire. The analysis of the gene-family distribution of targets by drug substance reveals that more than 50% of drugs target only four key gene families, from which almost a 30% correspond to the GPCRs superfamily [42]. This superfamily regulates the function of most cells in living organisms and is the focus of the work reported herein. The grouping of GPCRs into families or classes and these into types and subtypes based on sequence analysis may significantly contribute to helping drug design and to a better understanding of the molecular processes involved in receptor signalling both in normal and pathological conditions.

The challenge of managing the complexity of these types of data invites us to go one step further than traditional statistics and resort to intelligent pattern recognition approaches. In particular, statistical pattern recognition and machine learning methods bear the potential to both to scale well to large databases and to deal with non-trivial types of data. Sound statistical principles are essential to trust the evidence base built with any computational analysis of medical data [35]. Statistical machine learning methods are already establishing themselves in the more general field of bioinformatics [2].

The motivation of this thesis is the need for a robust probabilistic method capable of grouping and visualizing symbolic protein sequences, based on their structural and functional properties. As mentioned in [55], there is no biologically-relevant manner of representing the symbolic sequences describing proteins using real-valued vectors. This does not preclude the possibility of assessing the similarity between such sequences. Kernel methods can be used to this purpose if understood as similarity measures. Moreover, the visualization of high-dimensional protein sequence data can be the key exploratory tool for finding meaningful information that might be obscured by the intrinsic complexity of data [7].

In this thesis, the analysed data consist of sequences of GPCRs. These proteins were selected because an enormous amount of current pharmaceutical research is aimed at understanding their structure and function. They play an important role in human physiology and disease, but their three-dimensional structures remain mostly unsolved.

The following chapters report work on the grouping and visualization of GPCR protein sequences using a kernel variant of a non-linear model of the manifold learning family. A suitable kernel for this type of data is described. The visualization of the sequence data and the grouping results can be a useful tool in the quest for interpretability. The reported results reinforce the veracity of this statement.

Chapter 2

The biological problem

2.1 Current targets in the quest for new medicines

As stated in [42], there is a paradox in the fact that an industry such as pharma that spends yearly more than US $50 billion on R+D, has not been able to generate enough knowledge about the set of molecular targets that are the object of its products. That is why drug target discovery has, of late, received much attention from different areas of biochemistry-related drug research.

Arguably, drug research has contributed more to the progress of medicine in the past century than any other factor [12]. This is the result of advances in chemistry, pharmacology, and the clinical sciences. Molecular biology and genomics are now at the forefront of drug research. This has been exponentially amplified by developments in information, communication, and computation technologies. Genomics, proteomics, and the bioinformatics tools that support them can provide us with knowledge of suitable targets for medicines yet to be designed and, therefore, with a more proactive leverage on the process of drug design.

2.1.1 Receptors

Briefly, receptors can be defined in biochemistry as proteins to which signalling molecules may attach. They can be divided into two classes: the membrane-bound receptors and the soluble cytoplasmic or nuclear receptors. Receptors constitute the first step in the process of external signalling allowing the initiation of intracellular signalling cascades after specific ligand binding.

This thesis focuses on GPCRs, a particular set of membrane-bound receptors. GPCRs, as indicated by their name, signal through their interaction and subsequent activation of G proteins [22]. However, the functioning of these receptors appears more complex than was initially thought and additional accessory proteins play a role in the signal transduction concert. Proteins other than G proteins reported to interact with GPCRs and potentially responsible for G protein-independent GPCR signalling include β-arrestins, tyrosine kinases and PDZ-domain containing proteins [51]. Nevertheless, discussion on GPCR signalling pathways other than G proteins is beyond the scope of the present work and will not be included here.

2.1.2 GPCRs as pharmacological targets

GPCRs constitute the most abundant family of membrane-bound receptors and one of the largest in the whole human genome [46]. Analysis of the gene-family distribution of targets by drug substance reveals that more than 50% of drugs target only four key gene families, from which almost a 30% correspond to the GPCR superfamily. GPCRs regulate the function of most cells in living organisms.

GPCRs have been the subject of a vast research effort in the pharmaceutical industry due to their ubiquity and involvement in a broad spectrum of physiological functions. Moreover, drugs do not need to have the ability to cross the cell membrane to stimulate these receptors, thus increasing the size of the drug

discovery space and the possibility of success. Some examples of therapeutic indications for drugs acting on GPCRs are: antihistamines, anaesthetics, antidepressants, antipsychotics, anxiolytics, anti-ulcer, hypertension controllers, asthma, heart failure, Parkinson's, schizophrenia, migraines and cancer.

In this thesis, we have paid special attention to metabotropic glutamate receptors (mGlu receptors), a type of receptors belonging to GPCR family C that has generated a wealth of publications over the last few years (a search of the mGlu receptor string in *PubMed* [1] on 13th June 2011 produced 63 references for the year 2011), which shows that these receptors are very attractive as a pharmacological target for innovative drugs in neurological and psychiatric disorders.

2.2 GPCRs: Structure, function and classification

GPCRs consist of a single protein chain that crosses the membrane seven times [26]. For this reason they are also known as seven transmembrane (or 7TM) receptors. They constitute the most abundant family of membrane receptors and one of the largest in the whole human genome [46]. As mentioned previously, the name is derived from their association with heterotrimeric G proteins, which act as intermediary components, activating or inhibiting several intracellular effectors.

GPCRs were discovered in 1970 by Martin Robdell who determined the link between the activity of glucagon peptide and a molecule called guanosine triphosphate (GTP). At the same time, Alfred G. Gilman corroborated these results by finding the same trend in adrenergic receptors. The molecule responsible for the signal transduction was called G-protein [22]. These discoveries

[1] http://www.ncbi.nlm.nih.gov/pubmed/

allowed both researchers to share the Nobel Prize in 1994.

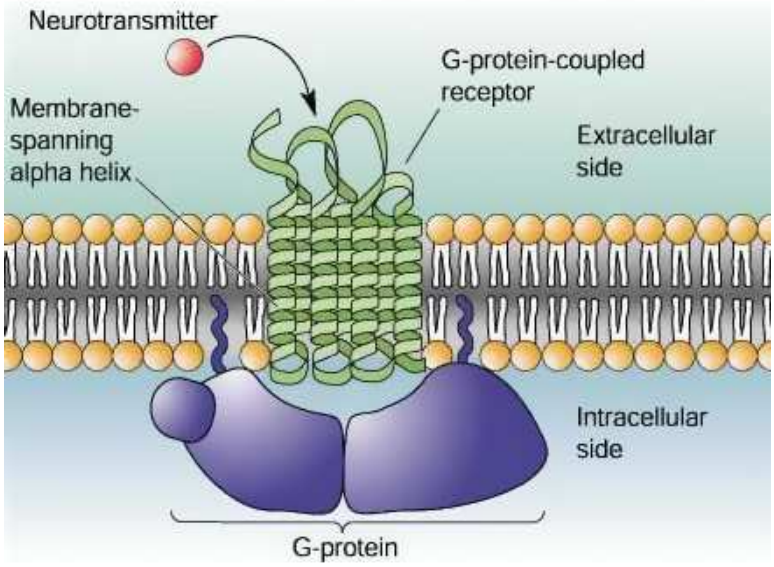


Figure 2.1: General structure of a GPCR protein

All GPCRs share a common general protein structure. The seven transmembrane helices are connected between them by three intracellular and three extracellular loops with varying lengths for each receptor subtype. The heptahelical transmembrane domain is largely hydrophobic, whereas the extracellular and intracellular segments, or loops, are generally hydrophilic. GPCRs have an extracellular amino terminus and an intracellular carboxyl terminus (See figure 2.1 courtesy of URL[2]). The most variable structures among the family of GPCRs are the carboxyl terminus, the intracellular loops and the amino terminus (See figure 2.2).

The ligands that bind and activate these receptors include light-sensitive compounds, various sensory signals (such as light and odors), pheromones, hormones, and neurotransmitters, and vary in size from small molecules to peptides, and to large proteins. Their stimulation leads to activation of specific

[2]http://www.csuci.edu/alzheimer/science/

G-proteins that transduce extracellular mediator messages to specific intracellular signalling pathways, producing particular cellular responses. In general, they can be thought as a communication tool between the cell and the outside world.

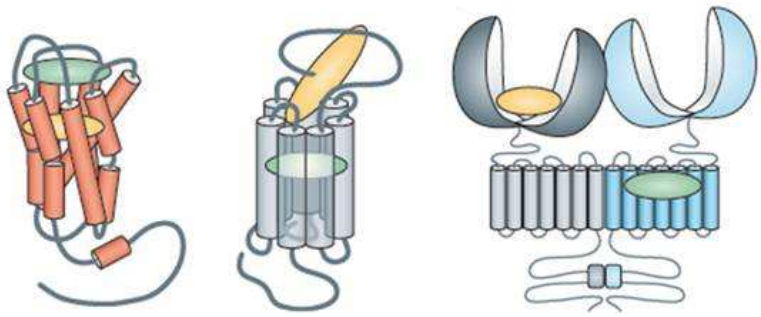


Figure 2.2: Structural diversity of GPCRs: family A, family B and family C [10]

The GPCRDB[3] [26], a database system for GPCRs, divides the GPCR superfamily into five major families (A to E) based on the ligand types, functions, and sequence similarities (summarized in table 2.1). Within the families, proteins are further divided into groups (types and subtypes) which bind common agents on the extracellular side of the membrane. The evolutionary relationship between groups is not known; they may have diverged from a common ancestor or be the result of convergent evolution, in which functional constraints lead to unrelated proteins from different organisms with the same design.

The sequences of different GPCR families are highly diverged from each other, except that they share one common structural feature, that is, they all have seven hydrophobic transmembrane regions. GPCRs within a family share common functions and more sequence similarities. Family A, the Rhodopsin like class, is by far the most populated GPCR family with more than 3,500 members in the database. Each family is further divided into groups, and so forth,

[3]http://www.gpcr.org/7tm/

Superfamily	Description
Family A	Receptors related to Rhodopsin and the beta2-adrenergic Receptors
Family B	Receptors related to the Calcitonin and PTH/PTHrP Receptors
Family C	Receptors related to the Metabotropic Receptors
Family D	Receptors related to the pheromone Receptors
Family E	Receptors related to the cAMP Receptors

Table 2.1: G-protein coupled receptor families

depending upon the common agents they bind to and sequence similarities.

While the identification of the function of GPCR sequences has a great importance in biomedical and pharmaceutical research, identifying and classifying this membrane protein superfamily is a difficult task due to the high levels of divergence observed among the GPCR family members. Therefore, it becomes important that there be a way to accurately and efficiently identify any new GPCRs from genomic data. As a consequence, this would benefit the pharmaceutical research and give us a better understanding of GPCR functions. GPCRs are used in this study due to their scientific importance, and also as an example of highly diverged protein families.

2.3 GPCR Family C

The family C of GPCRs have become an increasingly important target for new therapies, particularly in areas such as pain, anxiety, neurodegenerative disorders and as antispasmodics, but also potentially for the treatment of hyperthyroidism and osteoporosis.

In contrast to other GPCRs families, family C receptors are composed of three main structural domains, not including the C-terminal tail which can be very long and where a multitude of intracellular scaffolding and signaling

molecules bind. These domains are the Venus flytrap (VFT), which contains the agonist binding site (orthosteric site), the cysteine-rich domain (CRD) and the heptahelical domain (HD) involved in G-protein activation (See figure 2.3), and which contains potential allosteric sites to which synthetic allosteric ligands may bind. In addition, family C GPCRs have been shown to be constitutive dimers and therefore represent a good model for studying the functional relevance of GPCR dimerization [3].

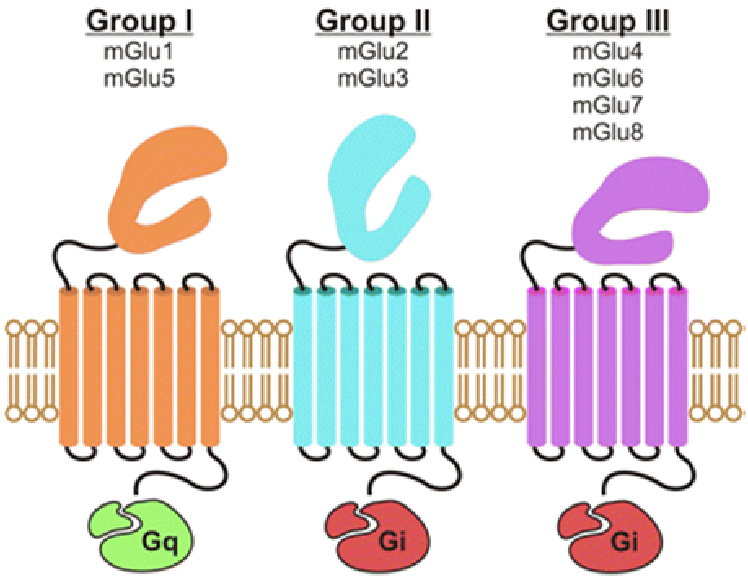


Figure 2.3: mGlu receptors are grouped into three families: group I, group II, and group III. [28]

Seven types of Family C, summarized in table 2.2, were investigated in this thesis, namely: Metabotropic glutamate, Calcium sensing, GABA-B, Vomeronasal, Pheromone, Odorant and Taste.

GPCR Family C	Description
Type 1	Metabotropic glutamate
Type 2	Calcium sensing
Type 4	GABA-B
Type 5	Vomeronasal
Type 6	Pheromone
Type 7	Odorant
Type 8	Taste

Table 2.2: GPCR Family C types

2.3.1 Metabotropic glutamate receptors

The metabotropic glutamate (mGlu) receptors, which belong to the first group of the GPCR Family C, are activated by glutamate, the major excitatory neurotransmitter in the central nervous system, and play important roles in regulating cell excitability and synaptic transmission. The mGlu receptors are widely distributed throughout the central nervous system, and a whole range of neurological and psychiatric disorders might be treated using drugs that act directly on these receptors.

There are eight types of mGlu receptors (eight genes encoding for mGlu1 to mGlu8 in humans) divided into three groups (See figure 2.3) according to structure, pharmacology and mechanism of signal transduction [50]:

- Group-I: mGlu1,mGlu5
- Group-II: mGlu2,mGlu3
- Group-III: mGlu4, mGlu6, mGlu7 and mGlu8

Like other components of the glutamatergic system, mGlu receptors also have a widespread distribution outside the CNS, including cells that do not have

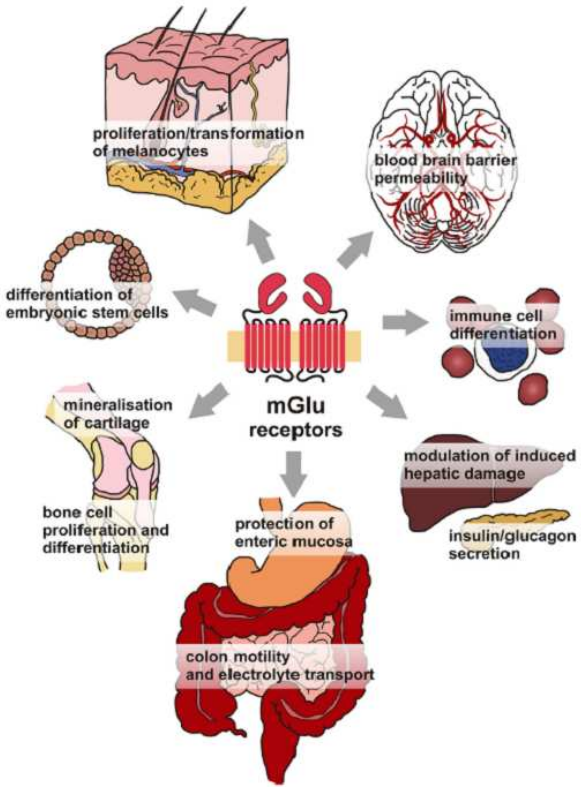


Figure 2.4: Summary of roles of mGlu receptors in peripheral tissues [28]

a neuronal phenotype (See figure 2.4). Analysis of the recent literature reveals an extraordinary potential, particularly for groups I and III, in the treatment of peripheral disorders of the most diverse nature, such as endocrine dysregulation, aberrant cell proliferation, and gastrointestinal disorders [28]. The significance of these findings is that pharmacological tools originally designed for mGlu receptors in the CNS may also be directed toward new disease targets in the periphery.

The wide diversity and heterogeneous distribution of mGlu subtypes provides

an opportunity for selectively targeting individual mGlu subtypes involved in only one or a limited number of CNS functions for the development of novel treatment strategies for psychiatric and neurological disorders. A large body of preclinical studies now suggests that ligands for specific mGlu subtypes have potential for the treatment of multiple CNS disorders, including depression, anxiety disorders, schizophrenia, pain syndromes, epilepsy, Alzheimer's disease, and Parkinson's disease, among others.

2.4 From the amino acid sequence to the structure and function of the protein

The function of the proteins depends directly on their 3D structure, which is embodied in their amino acid sequence. GPCRs are membrane proteins, and this environment makes their 3D structure difficult to unravel through nuclear magnetic resonance or X-ray crystallography. Knowledge about the three-dimensional structure of a GPCR is crucial for the understanding of its function and for the design of drugs. Modern molecular biology methods, though, make their amino acid sequences easy to acquire. The grouping of GPCRs into types and subtypes based on sequence analysis may significantly contribute to helping drug design and to a better understanding of the molecular processes involved in receptor signaling both in normal and pathological conditions [9].

Chapter 3

Analyzing Protein Sequences using Kernel Methods

The task of protein grouping, where proteins are specified by their amino acid sequences, aims to find biologically meaningful partitions of a given protein family. This might help the analyst to make inferences about key protein regions and residues both in the obtained groups and for the whole family. In order to get a better understanding of the functional role of the members of a protein family in biochemical processes, it is important to know the internal organization of the family and to detect key regions where interactions with other molecules may take place or which are essential to inform the three-dimensional structure of the protein.

The grouping of GPCRs into types and subtypes based on sequence analysis may significantly contribute to helping drug design and to a better understanding of the molecular processes involved in receptor signalling both in normal

and pathological conditions [28]. The importance of the GPCR as physiological agents and drug targets more than justifies our efforts in addressing this challenge.

In order to group GPCR sequences, we need a measure of similarity between them. Pattern recognition and machine learning techniques can help us in this task. Unsupervised data analysis using clustering algorithms provides a useful tool to explore data structures. Broadly speaking, the aim of clustering methods is that of grouping patterns on the basis of similarity (or dissimilarity) criteria, where the resulting groups or clusters are data subsets including similar patterns.

Unsupervised methods that were capable of providing simultaneous grouping and visualization of sequence data would be especially adequate for this type of problem, as visualization can help us to intuitively interpret the grouping and classification results by providing intuitive insights about the relationships between groups. The visualization of the high-dimensional GPCR sequences would considerably help to understand their global grouping structure.

The visualization of data clusters in low-dimensional spaces also becomes a dimensionality reduction task, for which linear and nonlinear modeling strategies can be used. Most of this strategies, though, have been designed for real-valued data. Needless to say, protein symbolic sequences of amino acids do not fit into this description, and alternative strategies are thus required. In this thesis, we resort to kernel methods. Over the last few years, several kernel methods for the visualization (and eventually clustering) of non-standard multivariate data have been proposed. The use of kernels allows mapping data implicitly into a high-dimensional space called feature space, in such a way that computing a linear partitioning in this feature space results in a corresponding nonlinear partitioning in the observed data space.

In the remaining of this chapter, we describe the basis of two of these methods that we consider to be representative of the current available choices in the

field. They should help to lay the conceptual foundations of the kernel manifold learning model used in this thesis, which is described in some detail in the next chapter.

3.1 Kernel Principal Component Analysis

Principal Component Analysis (PCA) [45] is an orthogonal transformation of the coordinate system in which we describe the observed multivariate data. The new coordinate system is obtained by projection onto the so-called principal axes of the data. The central idea of PCA is to achieve dimensionality reduction while retaining as much of the variation present in the data set as possible. Dimensionality reduction is achieved because a small number of principal components often suffices to account for most of the variance (structure) in the data.

This transformation yields a new set of variables or features (PCA can be classified as a feature extraction technique), known as principal components (PCs), which are uncorrelated, and which can be ordered so that the first few retain most of the variation present in all of the original variables. PCA takes an initial subset of these features and projects the observed data into the space it spans.

Data are effectively transformed by projecting them into the subspace spanned by the first k eigenvectors of the covariance matrix of the analyzed data set. The new coordinates are known as the principal coordinates with the eigenvectors referred to as the principal axes. Details of this technique can be found elsewhere [27].

Kernel PCA [53], or KPCA, is the application of PCA in a kernel-defined feature space making use of the dual representation. This method makes possible to detect nonlinear relations between variables in the data by embedding the data into a kernel-induced feature space, where linear relations can be found by

means of PCA. Also, KPCA can be seen as a way of inferring a low-dimensional explicit geometric feature space that best captures the structure of the data.

The projection of a new data point $\phi(x)$ onto the direction u_j in the feature space, is given by

$$P_{u_j}(\phi(x)) = u_j^{'}\phi(x) = \left\langle \sum_{i=1}^{l} \alpha_i^j \phi(x_i), \phi(x) \right\rangle \tag{3.1}$$

$$= \sum_{i=1}^{l} \alpha_i^j \langle \phi(x_i), \phi(x) \rangle = \sum_{i=1}^{l} \alpha_i^j K(x_i, x) \tag{3.2}$$

Hence, we will be able to project new data onto the eigenvectors in the feature space by performing an eigen-decomposition of the kernel matrix.

Let be U_k the subspace spanned by the first k eigenvectors in the feature space. Then, we can compute the k-dimensional vector projection of new data into this subspace as

$$P_{U_k}(\phi(x)) = \left(u_j' \phi_j(x)\right)_{j=1}^{k} = \left(\sum_{i=1}^{l} \alpha_i^j K(x_i, x) \right)_{j=1}^{k} \tag{3.3}$$

where $\alpha^j = \lambda_j^{-\frac{1}{2}} v_j$ is given in terms of the corresponding eigenvector λ_j and eigenvalue v_j of the kernel matrix. Equation 3.3 forms the basis of KPCA.

The critical question for assessing the performance of KPCA is the extent to which the projection captures new data drawn according to the same distribution as the training data. Therefore, we assess the stability of KPCA through the pattern function:

$$f(x) = \| P_{U_k}^{\perp}(\phi(x)) \|^2 = \| \phi(x) - P_{U_k}(\phi(x)) \|^2 =$$
$$\| \phi(x) \|^2 - \| P_{U_k}(\phi(x)) \|^2$$

That is, the squared norm of the orthogonal (residual) projection for the subspace U_k spanned by the first k eigenvectors. As always we wish the expected value of the pattern function to be small

$$E_X\left[f\left(x\right)\right] = E_X\left[\| P_{U_k}^{\perp}\left(\phi\left(x\right)\right) \|^2\right] \approx 0$$

Thus, capturing a high proportion of the data variance in an small number of dimensions is an indication that a reliable set of features has been detected and that the corresponding subspace will capture most of the variance of yet unobserved test data.

3.2 Kernel Self-Organizing Maps

KPCA provides a method according to which we can visualize GPCR sequences in a representation space (e.g. spanning only two PCs). Unfortunately, this visualization through projection is not accompanied by a grouping or clustering of the sequences. The Self-Organizing Map (SOM), popularly referred to as Kohonen network [31], [30] is a computational intelligence method for the visualization of high-dimensional data that also provides vector quantization and, in doing so, allows the partition of the data into clusters.

The SOM defines an topologically-ordered mapping that generates the projection of observed multivariate data items onto a regular, usually two-dimensional map. This map consists of a regular lattice of processing units, also called *neurons* (due to the original description of SOM as a bio-plausible model of cognitive processes). Each of these units is associated to a prototype vector in the observed data space, which can be considered as a representative example of a given subset of data cases. The map attempts to represent all the available data cases with optimal accuracy using a restricted set of prototypes. Each prototype could therefore be understood as a cluster representative.

The resulting map is meant to retain the topological order of the observed space, so that similar prototypes in the observed space are also close to each other in the visualization map.

In its standard form, the SOM algorithm distinguishes two stages: the competitive stage and the cooperative stage. In the former, the SOM *neuron* best matching a given data case is selected, while, in the latter, the coefficients (or *weights*) of the best-matching prototype (and to a lesser extent, those of its immediate lattice neighbors) are changed to become fractionally closer to that data case.

More formally, let $X = [x_1, x_2, ...x_d]^T \in R^d$ be the input vector. Assume a discrete lattice of units indexed with a index i. Each unit is associated to a corresponding weight vector (prototype) $W = [w_1, w_2, ...w_d]^T \in R^d$. Data case X_n is mapped to that unit whose weight vector is its nearest neighbour, from among all the weight vectors. This is called the best-matching unit(BMU) and is found as: $BMU_n = argmin_i \|X_n - W_i\|$

Thus, the training process of the SOM algorithm can be summarized as follows:

- For each observed data case, find out the nearest-neighbour (winner) from among the weight vectors associated to the map.
- Update the weights of the winner and all its neighbours according to some updating criterion.
- Iterate the process for all data cases (in an online or batch procedure) until some convergence criterion is met.

The SOM model, though, has some limitations due to its heuristic nature. In summary:

- Different runs of the SOM algorithm with different initializations yield different results.
- The selection of its parameters (e.g., learning rate, or neighbourhood function type or size) has no theoretical basis.

- There is no guarantee of error convergence for the training procedure. Neighbourhood preservation is not guaranteed either.
- There is no theoretical basis for complexity control (regularization and overfitting)

Furthermore, the Euclidean distance used to describe similarity in the standard SOM model is not adequate for the analysis of non-real-valued data such as symbolic protein sequences.

Recently, a kernel version of the SOM, namely the Kernel Self-Organizing Map, or KSOM, was proposed by MacDonald and Fyfe [37]. It can be understood as a kernelization of the k-means clustering algorithm, but with added neighbourhood learning. More precisely, a kernel function is applied to transform the input (observed data) into a high-dimensional feature space, thus transforming the distance metric to nonlinear and adding more flexibility in the vector-quantization process in order to better capture the data structure [33]. Each data case x is mapped to the feature space via a nonlinear function $\phi(x)$. In principle each mean can be described as a weighted sum of the observations in the feature space,

$$m_i = \sum_n \gamma_{i,n} \phi(x_n)$$

where $\{\gamma_{i,n}\}$ are the constructing coefficients. The algorithm then selects a mean or assigns a data case with the minimum distance between the mapped point and the mean,

$$\|\phi(x) - m_i\|^2 = \| \phi(x) - \sum_n \gamma_{i,n} \phi(x_n) \|^2 = \tag{3.4}$$

$$K(x,x) - 2\sum \gamma_{i,n} K(x, x_n) + \sum_{n,m} \gamma_{m,n} K(x_n, x_m) \tag{3.5}$$

The update of the mean is based on an update expression similar to that of the SOM:

$$m_i(t+1) = m_i(t) + \Lambda[\phi(x) - m_i(t)] \tag{3.6}$$

where Λ is the normalized winning frequency of the i-th mean, defined as:

$$\Lambda = \frac{\xi_{i(x),j}}{\sum_{n=1}^{t+1} \xi_{i,n}} \tag{3.7}$$

and ξ is the winning counter and is often defined as a Gaussian function between the indexes of the two neurons. As the mapping function ϕ is not known, the updating rule 3.6 is further elaborated and leads to the following updating rules for the constructing coefficients of the means [37]:

$$\gamma_{i,n}(t+1) = \begin{cases} \gamma_{i,n}(t)(1-\xi), & for\ n \neq t+1 \\ \xi, & for\ n = t+1 \end{cases}$$

Note that these constructing coefficients, $\gamma_{i,n}$, together with the kernel function, effectively define the kernel SOM in the feature space. The winner selection, i.e. 3.4, operates on these coefficients and the kernel function. No explicit mapping function ϕ is required. The exact means or neuron weights m_i, are not required [62].

There is an alternative direct way to kernelize the SOM by mapping the data points and neuron weights, both defined in the input space, to a feature space; this is followed by applying standard SOM in the mapped dot-product space. The winning rules of this second type of KSOM have been proposed as follows, either in the input space [44], $v = arg\min_i \|x - m_i\|$ or in the feature space [1], $v = arg\min_i \|\phi(x) - \phi(m_i)\|$

These two rules are equivalent for certain kernels, such as the Gaussian. The weight update rule proposed in [1] is:

$$m_i(t+1) = m_i(t) + \alpha(t)\,\eta(v(x), i)\,\nabla J(x, m_i) \tag{3.8}$$

where $\nabla J(x, m_i) = \|\phi(x) - \phi(m_i)\|^2$ is the distance function in the feature space or the proposed instantaneous or sample objective function. Also, $\alpha(t)$ and $\eta(v(x), i)$ are, in turn, the learning rate and neighbourhood function.

Note that,

$$J(x, m_i) = \|\phi(x) - \phi(m_i)\|^2 = K(x, x) + K(m_i, m_i) - 2K(x, m_i)$$

and,

$$\nabla J(x, m_i) = \frac{\partial K(m_i, m_i)}{\partial m_i} - 2\frac{\partial K(x, m_i)}{\partial m_i}$$

Therefore this kernel SOM can also be operated entirely in the feature space with the kernel function. As the weights of the neurons are defined in the input space, they can be explicitly resolved.

The standard SOM minimizes the following energy function [32], [25]:

$$E = \sum_i \int_{V_i} \sum_j \eta(i, j) \| x - m_j\|^2 p(x)\, dx$$

where V_i is the Voronoi region of neuron i.

The extension of this energy function in the feature space is:

$$E_F = \sum_i \int_{V_i} \sum_j \eta(i, j) \| \phi(x) - \phi(m_j)\|^2 p(x)\, dx$$

The KSOM can be seen as a result of directly minimizing this transformed energy. Using the sample gradient on $\eta(v(x), j) \| \phi(x) - \phi(m_j)\|^2$, we obtain:

$$\frac{\partial \hat{E_F}}{\partial m_i} = \frac{\partial}{\partial m_j} \sum_j \eta(v(x), j) \| \phi(x) - \phi(m_j)\|^2 = -2\eta(v(x), i) \nabla J(x, m_j),$$

which leads to the same weight update expression for the KSOM as in equation 3.8.

Although KSOM makes the standard Kohonen map much more flexible, it still inherits the limitations of SOM outlined above. The analysis of GPCR

sequences would benefit from a model with solid grounds on probability theory that might benefit from the automatic optimization of all its parameters. One such kernel model of the manifold learning family is proposed and applied to the analysis of GPCR sequences in the following chapter.

Chapter 4

Grouping and visualization of GPCRs using Kernel GTM

As stated in previous chapters, the grouping of GPCRs into types and subtypes according to the amino acid symbolic sequences that describe them can provide useful insights for the design of targeted pharmacological drugs. The understanding of this grouping structure would benefit from its low dimensional visual representation. Unfortunately, standard clustering techniques are of little use for the grouping of symbolic sequences. Kernel methods can bypass this limitation through the definition of kernels that appropriately describe similarity and dissimilarity between sequences. Therefore, our target is the definition of a kernel method for the simultaneous grouping and visualization of symbolic sequences. We also want this method to be grounded on sound theoretical foundations, and to have the ability to estimate the most adequate values for its constituting parameters.

4.1 Kernel Generative Topographic Mapping

The Generative Topographic Mapping (GTM) [4] is a non-linear latent variable model of the manifold learning family, with sound foundations in probability theory. It performs simultaneous clustering and visualization of the observed data, through a non-linear and topology-preserving mapping from a visualization latent space in in $\mathbb{R}^{\mathbb{L}}$ (with L being usually 1 or 2 for visualization purposes) onto the $\mathbb{R}^{\mathbb{D}}$ space in which the observed data reside. The mapping that generates the embedded manifold takes the form:

$$y = W\phi(u), \tag{4.1}$$

where u is an L-dimensional point in latent space, W is the matrix that generates the mapping, and ϕ consists of S basis functions ϕ_s(radially symmetric Gaussians in the standard model for continuous data). To achieve computational tractability, the prior distribution of u in latent space is constrained to form a uniform discrete grid of M centres, analogous to the layout of the SOM units, in the form of a sum of delta functions:

$$p\left(u\right) = \frac{1}{M}\sum_{m=1}^{M}\delta\left(u - u_m\right), \tag{4.2}$$

where M is the number of nodes in the grid.

This way defined, the GTM can also be understood as a special case of a Gaussian mixture model that is adapted to provide high-dimensional data visualization. Each component m in the mixture defines the probability of an observable data point x given a latent point u_m and model:

$$p(x \mid u_m, \Theta) = \left(\frac{\beta}{2\pi}\right)^{\frac{D}{2}} \exp\left\{-\frac{\beta}{2}\left\|x - y_m\right\|^2\right\} \tag{4.3}$$

where D is the dimensionality of the data space, and $y_m = W\phi\left(u_m\right)$.

The set of adaptive parameters Θ is constituted by W and the common inverse variance β. A density model in data space is therefore generated for

each component m of the mixture, which, assuming that the observed data set X consists of N independent, identically distributed (i.i.d.) data points x_n, leads to the definition of a likelihood in the form:

$$\mathcal{L}(W,\beta) = \prod_{n=1}^{N} \frac{1}{M} \sum_{m=1}^{M} p(x_n \mid u_m, W, \beta) \tag{4.4}$$

However, it is more convenient to work with the log-likelihood function:

$$L(W,\beta) = \sum_{n=1}^{N} \ln \left\{ \frac{1}{M} \sum_{m=1}^{M} p(x_n \mid u_m, W, \beta) \right\} \tag{4.5}$$

The adaptive parameters of the model are usually optimized by Maximum Likelihood (ML) using the Expectation-Maximization (EM) algorithm [11]. In the E-step, we use the current values of the parameters W and β to evaluate the posterior probability, or *responsibility*, which each component m takes for every data point x_n, which, using Bayes' theorem, is given by

$$R_{nm} \equiv p(m \mid x_n) = \frac{p(x_n \mid m)}{\sum_j p(x_n \mid j)}, \tag{4.6}$$

in which the prior probabilities $P(m) = \frac{1}{K}$ have cancelled between numerator and denominator. Using 4.3, we can rewrite this in the form

$$R_{nm} = \frac{exp\left\{-\frac{\beta}{2} \parallel x_n - y_m \parallel^2\right\}}{\sum_m exp\left\{-\frac{\beta}{2} \parallel x_n - y_m \parallel^2\right\}} \tag{4.7}$$

Then in the M-step we use the responsibilities to re-estimate the weight matrix W by solving the following system of linear equations:

$$\left(\Phi^T G \Phi\right) W_{new}^T = \Phi^T R X, \tag{4.8}$$

which follow by maximization of the expected complete-data log likelihood. In 4.8, Φ is a $K \times M$ matrix with elements $\Phi_{mj} = \Phi_j(u_m)$, X is an $N \times D$ matrix with elements x_{nk}, R is a $K \times N$ matrix with elements R_{nm}, and G is a $K \times K$ diagonal matrix with elements. The inverse variance parameter is

also re-estimated in the M-step using $G_{mm} = \sum_n R_{nm}$. The inverse variance parameter is also re-estimated in the M-step using:

$$\frac{1}{\beta_{new}} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{m=1}^{M} R_{nm} \parallel W_{new}\phi\left(u_m\right) - x_n \parallel^2 \tag{4.9}$$

We can initialize the parameters W so that the GTM model initially approximates PCA. To do this, we first evaluate the data covariance matrix and obtain the eigenvectors corresponding to the q largest eigenvalues, and then we determine W by minimizing the sum-of-squares error between the projections of the latent points into data space by the GTM model and the corresponding projections obtained from PCA. The value of β^{-1} is initialized to be the larger of either the $q+1$ eigenvalue from PCA (representing the variance of the data away from the PCA sub-space) or the square of half of the grid spacing of the PCA-projected latent points in data space.

The main advantage of the GTM over the SOM model is that the former generates a density distribution in the input data space so that the model can be described and developed within a principled probabilistic framework. An example of development of the GTM is the use of a Bayesian approach to automatic regularization and smoothing of the resulting mapping. As part of this process, the GTM learning parameters calculation is grounded in a sound theoretical basis. The GTM also provides the well-defined objective function of equation 4.5, whereas the SOM training does not involve the minimisation of any error function; its maximisation using either standard techniques for non-linear optimisation or the EM-algorithm has been proved to converge, unlike in the case of the SOM.

4.1.1 Kernelization of the GTM

Kernelization is a method originally defined for Support Vector Machines (SVM) which could be used to develop generalizations of any algorithm that could be

cast in dot product terms. The idea is that a method formulated in terms of kernels can use the one that best suits the problem and data type at hand. With this purpose, we define kernel-GTM (KGTM). It takes advantage of the original GTM functionalities to achieve clustering and visualization of a wider variety of data types [41]. Moreover, GTM lacks the ability to handle more structured data, such as strings.

4.1.2 The KGTM model and its application to sequence analysis

Let us consider the problem of embedding GPCR sequences in a high-dimensional space in such a way that their relative position in that space reflects their similarity and that the inner product between their images can be computed efficiently. The first decision to be made is what similarity notion should be reflected in the embedding, or, in other words, what features of the sequences are informative for the task at hand.

The meaning of similarity in biological applications is related to both functional similarity and symbolic sequence similarity, the latter measured by the number of insertions, deletions and symbol replacements. Measuring sequence similarity should therefore give a good indication about the functional similarity that bioinformatics researchers would like to capture.

The similarity between two sequences is usually evaluated by first aligning the sequences (or parts of them) and then deciding whether their alignment is more likely to have occurred because the sequences are related or just by chance.

When two sequences are compared, the basic mutational processes under consideration are *substitutions*, which change residues in a sequence, and *insertions* and *deletions*, which add or remove amino acids in the sequence. Insertions and deletions are together referred to as *gaps*. Natural selection has an effect on this process by screening the mutations, so that some types of changes remain

throughout evolution and appear more often than others [14]. In order to have some control over the number of gaps, their size, position, etc., usually gap penalties are introduced. Then, the score, used to judge the correctness of the alignment, is modified accordingly to allow the number of gaps to be limited. The total similarity score assigned to an alignment will be a sum of terms for each aligned pair of residues, plus terms for each gap. In a probabilistic interpretation, this corresponds to the logarithm of the relative likelihood that the sequences are related, compared to being unrelated. Informally, identities and conservative substitutions are expected to be more likely in alignments than appearing by chance and, therefore contribute positively to the similarity score. On the contrary, non-conservative changes are expected to be observed less frequently in real alignments than expected by chance, and so they contribute negatively.

In order to gauge similarity for each aligned residue pair, we will derive substitution scores from our probabilistic model. The scores can be arranged in a matrix. For the protein sequences analyzed in this thesis, consisting on arrays of symbols from a 20 amino acid *alphabet*, a 20×20 matrix can be calculated, known as *score matrix* or *substitution matrix*.

A kernel function can be thought of as a measure of similarity between sequences. Different kernels correspond to different notions of similarity, and can lead to discriminative functions with different performance. The kernel function designed to analyze GPCRs with KGTM is a variation on that described in [41], based on the mutations and gaps between sequences:

$$K\left(x, x'\right) = exp\left\{\nu \frac{\pi\left(x, x'\right)}{\sqrt{\pi\left(x, x\right)\pi\left(x', x'\right)}}\right\} \tag{4.10}$$

where x and x' are two sequences and ν is a prefixed parameter; $\pi\left(.\right)$ is a score function commonly used in bioinformatics and expressed as: $\pi\left(x, x'\right) = \sum_r s\left(x_r, x'_r\right) - \gamma$, where x_r and x'_r are the r^{th} residue in the sequences. The

value of $s(x_r, x'_r)$ can be found in a mutation matrix [14] and γ is a gap penalty (usually the number of gaps in sequences). A normalization factor, defined as the geometric mean of the maximum scores for each of the sequences, is used in the kernel function instead of the sum used in [41].

4.1.3 The KGTM algorithm

The kernel trick allows the observed data X to be implicitly mapped into a high-dimensional feature space H via a nonlinear function: $x \longmapsto \psi(x)$. A similarity measure can then be defined from the dot product in space H as follows:

$$K\left(x, x^{'}\right) = \left\langle \psi(x), \psi\left(x^{'}\right) \right\rangle \tag{4.11}$$

K is a kernel function that should satisfy Mercer's condition [54]. It allows us to deal with learning algorithms using linear algebra and analytic geometry. In general, this method deals with data in the high-dimensional dot product space H, usually known as feature space.

The use of kernel trick avoids the explicit estimation of ψ, whose dimension is usually unknown (or even infinite).

The kernelization of GTM can be implemented by redefining equation 4.3 in feature space as:

$$p(\psi(x) \mid u_m, \Theta) = \left(\frac{\beta}{2\pi}\right)^{\frac{D}{2}} \exp\left\{-\frac{\beta}{2}\|\psi(x) - y_m\|^2\right\} \tag{4.12}$$

Note that the prototypes y_m are now defined in the feature space and not in data space, as originally. In most cases, the term $\|\psi(x) - y_m\|^2$ cannot be directly evaluated, given that the function $\psi(\cdot)$ is usually unknown. However, this term can be also expressed as follows:

$$\|\psi(x) - y_m\|^2 = \langle \psi(x), \psi(x) \rangle + \langle y_m, y_m \rangle - 2\langle \psi(x), y_m \rangle \tag{4.13}$$

Here, we assume that, as in KPCA, y_m can be expanded on the training data in the feature space. That is, $y_m = \mathbf{\Psi} w_m$, where $\mathbf{\Psi}$ is a $D \times N$ -matrix

of vector columns $\mathbf{\Psi}(x_n)$, $n = 1..N$, and w_m a weight vector. With the aim of preserving the topology, we correlate the weight vector to the latent space by $w_m = \Lambda\phi_m$, where Λ is an adaptive weight matrix and $\phi_m = \phi(u_m)$ is the set of radial basis functions typically used by GTM. Therefore, equation 4.13 becomes:

$$\|\psi(x) - y_m\|^2 = J_{mn} = K_{nn} + (\Lambda\phi_m)^T \mathbf{K}\Lambda\phi_m - 2k_n\Lambda\phi_m \tag{4.14}$$

where $\mathbf{K}$ is a kernel matrix with elements $K_{nn'} = \langle\psi(x_n), \psi(x_{n'})\rangle$, and row vectors k_n . Thereby J_{mn} is expressed in terms of the kernel matrix, making the definition of function $\psi(\cdot)$ unnecessary. The adaptive parameters of the model are now Λ and β , which can be optimized by ML using EM, as in GTM. The likelihood of the model is formulated as follows:

$$\mathcal{L}(\Lambda, \beta) = \prod_{n=1}^{N} \frac{1}{M} \sum_{m=1}^{M} p(\psi(x_n) \mid u_m, \Lambda, \beta) \tag{4.15}$$

Following the usual EM algorithm, we are specially interested in one of the results of the expectation step of EM, namely the estimation of the posterior distribution $R_{mn} = p(u_m \mid \psi(x_n), \Lambda, \beta)$, defined as:

$$R_{mn} = \frac{p(\psi(x_n) \mid u_m, \Lambda, \beta)}{\sum_{m'=1}^{M} p(\psi(x_n) \mid u_{m'}, \Lambda, \beta)} \tag{4.16}$$

R_{mn} measures the degree of responsibility (probability) of a point u_m in the latent space for the generation of a $\psi(x_n)$ GPCR data subsequence. In turn, each R_{mn} is an element of a $M \times N$ responsibility matrix R.

In the maximization step we use equation 4.15 as the optimization function to determine the parameters Λ and β, which results in the following expressions:

$$\Lambda^T = \left(\Phi^T G \Phi\right)^{-1} \Phi^T R \tag{4.17}$$

$$\frac{1}{\beta} = \frac{1}{ND} \sum_{n=1}^{N} \sum_{m=1}^{M} R_{mn} J_{mn} \tag{4.18}$$

The initial values for the parameters of KGTM are selected using KPCA (a procedure which is inspired in the PCA-based initialization of parameters for the standard GTM).

4.2 The GPCR dataset

The dataset analyzed in this thesis and used to assess the performance of KGTM in the grouping and visualization of GPCR consists of 232 amino acid sequences obtained from de public GPCR database and information system GPCRDB [26], corresponding to seven types (type 3 was not analyzed as it was not included in the GPCRDB database) belonging to the family C.

The GPCRDB stores three kinds of experimental data: sequences, mutation data and ligand binding data. In this database, sequences had been already downloaded previously aligned.

Each position in a sequence is called a residue, which in turn may be one of 20 possible amino acids (See table 4.1). Each amino acid has a standard one-letter code, and a sequence is therefore represented by a combination of these letters.

The number of residues by sequence (that is, the data dimensionality) in the analyzed dataset is 253. Table 4.1 shows the nomenclature for each amino acid, and the codes for the complete dataset are shown in table A.1. They are listed so that any other researcher interested in replicating our results could do so in an informed way. Two examples of sequences are shown in table 4.2 in FASTA format [34].

Amino acid name	Letter	Amino acid name	Letter
Alanine	A	Leucine	L
Arginine	R	Lysine	K
Asparagine	N	Methionine	M
Aspartate	D	Phenylalanine	F
Cysteine	C	Proline	P
Glutamate	E	Serine	S
Glutamine	Q	Threonine	T
Glycine	G	Tryptophan	W
Histidine	H	Tyrosine	Y
Isoleucine	I	Valine	V

Table 4.1: List of the 20 possible amino acids that can have a residue.

Header	Sequence
$ts1r3\_mouse$	RPKFLAWGEPVVLSLLLLLCLVLGLALAALGLSLVQA SGGSQFCFGLICLGLFCLSVLFPGRPSSASCLAQQPM AHLPLTGCLSTLFLQAAETFVESELPLSWNWLCSYLR GLWAWLVVLLATFVEAALCAWYLIAFPPEVVTDWSLP TEVLEHCHVRSLGLVHITNAMLAFLCFLGTFLVQSQP YNRARGLTFAMLAYFITWVSFVPLLANVQVAYCALGI LVTFHLPKCYVLLWLPKLNTQEFFLGRNAKK
$q7pfp4\_anoga$	−FAFYTVVILSLIGIGISVLFLGLNLRF− − − − −ST ITVCGCMLVYTATILLGLDHSTL− − − − −−STICMRIY FLSAGFSLAFGSMFAKTFRVYRIFTH− − − − −LISVIG ALLLVDAFVVSFWMAAD− −C− − −WLG MLYAYKGLLLLVGVYMAWQTRNVK−−NDSQ YIGISVYSV VITSASVVVLANLLYERIITAG FVLISTTATLCLLFLPKI− − − − − − − − − − − − − −

Table 4.2: Two sequences from the dataset. The first column represents the ID or header of the sequence and the second one represents the inner sequence. The gaps are represented by '−'.

4.3 KGTM grouping and visualization of GPCRs

The KGTM is a fully unsupervised model, that is, it not use any GPCR type or subtype label, even if known, as part of the data modelling process. The labelling of individual sequences is accomplished *a posteriori*, in order to assess the sequence grouping and visualization results.

The visualization of the GPCR sequences in the low-dimensional, latent representation space of the KTGM is accomplished through the use of the mode-

projection, defined as:

$$m_{mode} = arg\max_{m} R_{mn} \tag{4.19}$$

where R_{mn} is the responsibility (probability) of a point u_m in the latent space of KGTM for the generation of of the feature space-transformed sequence. Note that using equation 4.19 entails selecting that latent point for which the responsibility for a given sequence is maximum. This could be understood as a summary measure (*winner takes all*) according to which we can assign a given sequence to a specific latent point (or *cluster representative*). Even if only this summary measure is used to simplify the visualization procedure, we should not forget that there is still a non-zero probability of a given sequence belonging to any other cluster (latent point in the KGTM grid).

The basic visualization results using KGTM for the analyzed GPCR dataset are shown in figure 4.1. Seven types of family C have been modelled and and their sequences assigned to clusters in the KGTM representation map according to equation 4.19.

Overall, a clear separation between many of the GPCR types can be observed. Many receptor types occupy a rather differentiated area on the map, showing little overlapping. Thus, mGlu receptors (type 1), GABA-B (4), and Taste (8) are clearly differentiated from the rest of types, the latter showing significant overlapping between them. This overlapping can be, in part, pharmacologically explained. Odorant, Vomeronasal and Pheromone types, which present a high degree of overlapping, are related with the sense of smell. On the other hand, the obtained mixing of these receptors with the Calcium Sensing receptor is consistent with the branches distribution observed in published phylogenetic tree analyses [47] (see below for comparison of KGTM and phylogenetic tree methods of classification). Finally, the mixing between some mGlu receptors and receptors associated to smell sense requires further analyses (a

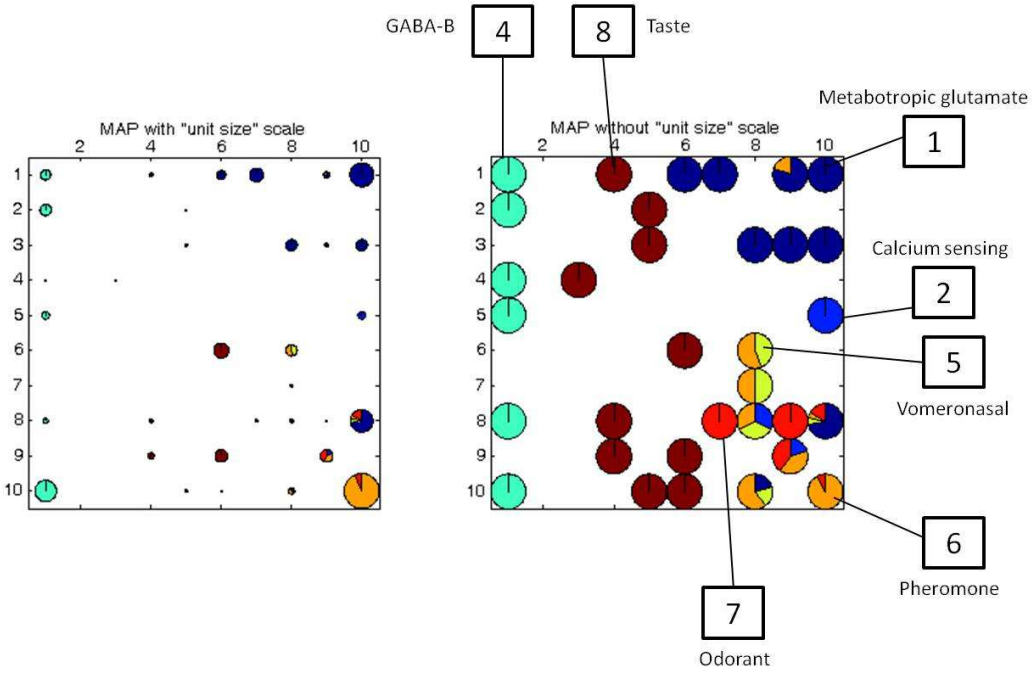


Figure 4.1: KGTM-based data visualization on a 10×10 representation map using the mode projection as described in the text. Left) Pie charts represent individual latent points and their size is proportional to the ratio of sequences assigned to them. Each portion of a chart corresponds to the percentage of sequences belonging to each type. Right) The same map without sequence ratio size scaling, for better visualization. Labels: 1: Metabotropic glutamate, 2: Calcium sensing, 4: GABA-B, 5: Vomeronasal, 6: Pheromone, 7: Odorant, 8: Taste.

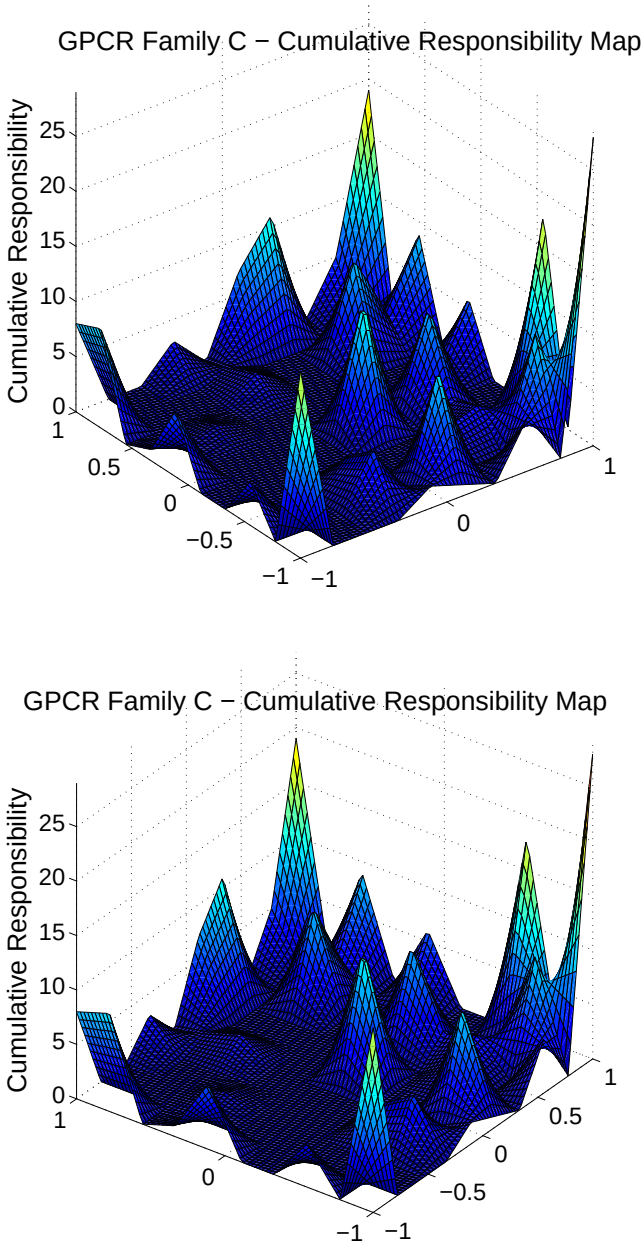
The mode-projection is an intuitive form of visualization that, as previously mentioned, sacrifices detail in favour of clarity. By using only the maximum of the responsibilities in R, though, it disposes of much of the rich information that might be contained in this matrix of probabilities. There are different ways of visually representing this information. One of them is the display of full maps of probability R_i, for a given sequence i.

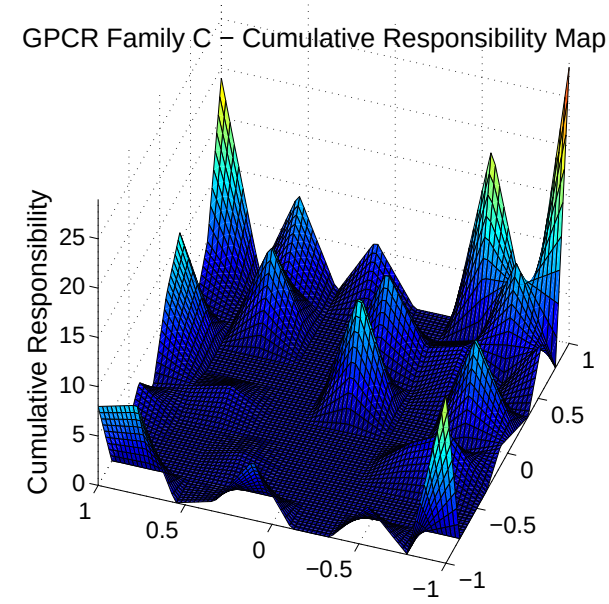
Sequences clearly ascribed to a type are likely to have their responsibilities

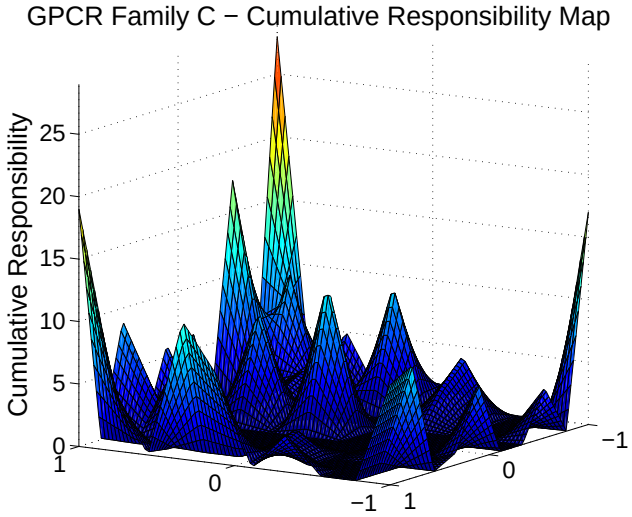
concentrated in only a few modes (latent points), whereas the probabilities of sequences without clear type ascription may be more evenly spread across the map.

We may be also interested in the responsibilities of all sequences of a given type at once. In this case, we would aim to assess if each type has its responsibilities located in a well-defined area of the map or not. For this, we use the cumulative responsibility of the sequences that belong to a given type c, which is defined as a vector $CR_c = \sum_{\{n \in c\}} R_{mn}$, for $m = \{1, .., M\}$.

Figure 4.2: Visualization of the global CR (on the vertical axis) of the data set on the representation map. For better appreciation, several viewpoints of the map are provided.





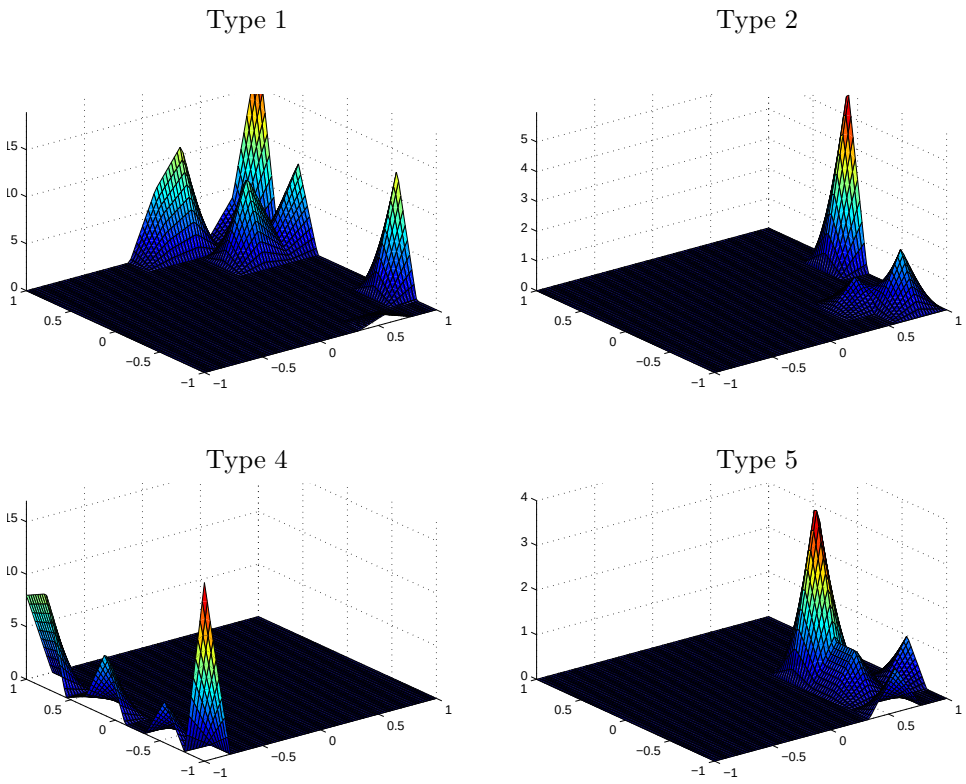


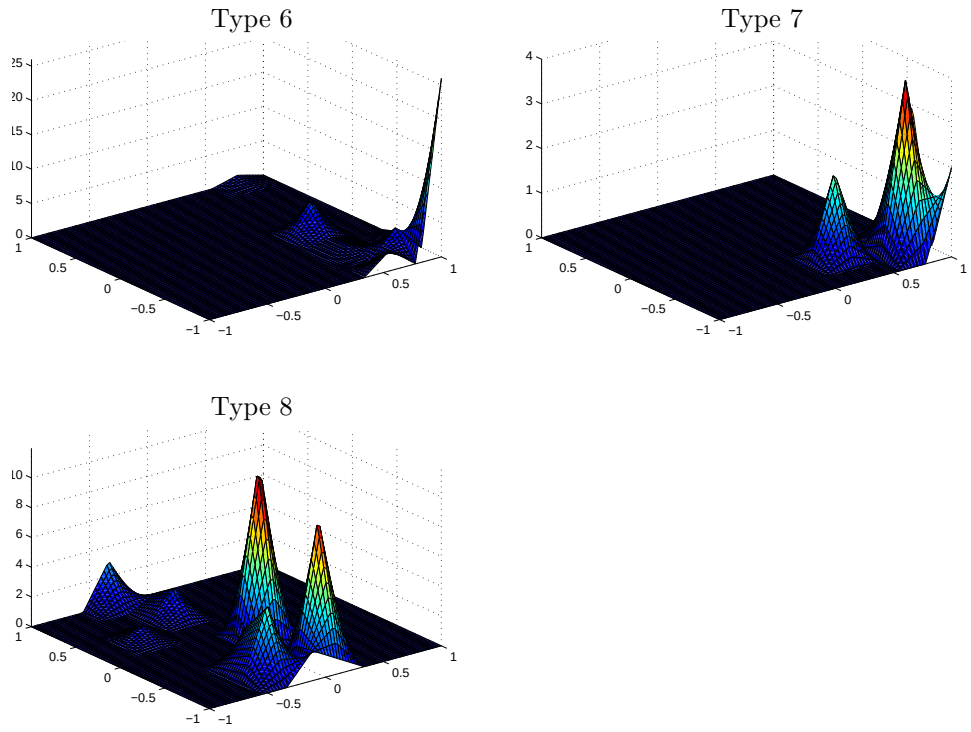
This takes us to the possibility of displaying the cumulative responsibility of all sequences in the database. With this map of probability, the existence

of CR peaks and valleys can be explored. The latter are likely to define the boundaries between types.

The global CR is displayed in figure 4.2, whereas figure 4.3 provides the visualization of the CR_c for the seven analysed types of the family C. Consistent with the type specific representations in figure 4.3, several local maxima are shown to correspond to each type, which could be an indication of heterogeneity within the types. Some deep valleys of probability can be seen in the central parts of the map in figure 4.2, drawing clear boundaries between the types represented in the periphery of the map and those around its center. Some amongst the latter are the ones with a higher level of mixing and would merit further investigation.

Figure 4.3: CR_c representation maps for all GPCR family C types. Labels: 1: Metabotropic glutamate, 2: Calcium sensing, 4: GABA-B, 5: Vomeronasal, 6: Pheromone, 7: Odorant, 8: Taste. Type 1 (Metabotropic glutamate), the most populated, is well-defined on the top-right corner of the map; type 4 (GABA-B), also isolated and unmixed in the left hand-side of the map; type 6 (Pheromone), strongly focused on the bottom right corner of the map, but partially overlapping with right: type 7 (Odorant). The layout corresponds to that of figure 4.1, although with its viewpoint slightly displaced to the left, to provide some perspective.





Our results are consistent with early classification studies using other techniques such as Hidden Markov Models [48], thereby validating the present methodology. Importantly, the KGTM mapping reveals mixing between some receptor types, suggesting its possible applicability to the study of heterodimerization [3] between receptors. Receptor heterodimerization has been confirmed experimentally for a number of receptors. This finding paves the way for new strategies in drug discovery research providing a conceptual framework for the rational combination of drugs. KGTM may help in the exploration of receptors susceptible of heterodimerization and thus be useful in the quest of more potent and safer drugs.

4.3.1 Zooming into the mGlu receptor GPCR subtype

As described in subsection 2.3.1, the mGlu receptors, widely distributed throughout the CNS, play a relevant role in the regulation of cell excitability and synaptic transmission.

They are divided into three groups (I, II, III) including eight subtypes: Group-I: mGlu1, mGlu5; Group-II: mGlu2, mGlu3; Group-III: mGlu4, mGlu6, mGlu7 and mGlu8.

The mode projection-based KGTM visualization of the mGlu receptors in our dataset is displayed in figure 4.4. It reveals the distribution of the eight different subtypes extracted from the GPCRDB dataset.

It is worth mentioning that in the primary data set the mGlu7 subtype was absent. Instead, a new subtype denoted as *mGluLike* was present. It may be assumed that the *mGluLike* subtype includes receptors that are classified as mGlu receptors by the GPCRDB program but without a fully true genetic adscription.

Strikingly, KGTM separates quite well each of the eight subtypes within the mGlu receptor type. Further detail of the mapped location of each subtype can be appreciated in the display of figure 4.4.

It is worth mentioning that the plot of mGlu subtypes displayed in figure 4.4 has been done on the KGTM model obtained previously for family C. In other words, the KGTM model was not trained again on the mGlu subset; instead, the other types were made "silent" and sequences were labelled accordingly with their mGlu receptor subtype identity.

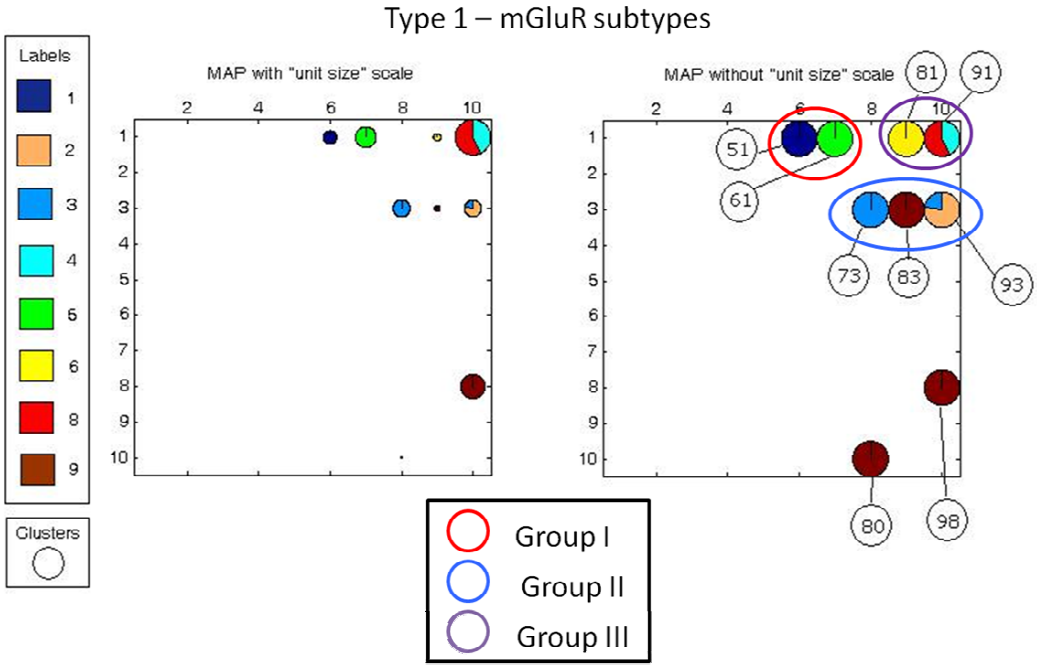


Figure 4.4: Mode projection of the type 1 mGlu receptor subtypes. Labels: 1: mGlu1, 2: mGlu2, 3: mGlu3, 4: mGlu4, 5: mGlu5, 6: mGlu6, 8: mGlu8, 9: *mGluLike*. The analysed dataset has no mGlu7 subtype cases. There is a visible separation of the subtypes in three main groups, according to the amino acid sequence similarity, agonist pharmacology and the signal transduction pathways to which they couple: group I (mGlu1, mGlu5), group II (mGlu2, mGlu3, *mGluLike*) and group III (mGlu4, mGlu6, mGlu8)

According to this visualization, the mGlu receptor sequences of subtype 9 corresponding to *mGluLike*, assigned to cluster 83 are very homogeneous. They include the subtypes *mGluLike2* and *mGluLike3* and are well-located between *mGlu2* and *mGlu3*. On the other hand, the *mGluLike* groups assigned to clusters 98 and 80 are quite far from Type 1- *mGlu receptors* but very close to the Types 5 *(Vomeronasal)*, 6 *(Pheromone)* and 7 *(Odorant)* (See figure B.8 in

Appendix B for complete detail), taking into account their neighbourhood. This suggests that some GPCRDB assignments of *mGluLike* receptors to the mGlu group might be incorrect, and that they might in fact be smell sense receptors. This is only a hypothesis and would require further testing.

4.4 KGTM and phylogenetic tree representations of GPCR

4.4.1 From protein sequences to phylogenetic trees

Generally speaking, a phylogenetic tree is a dendogram-like graphical representation of the evolutionary relationship between taxonomic groups which share a set of homologous characters. In biology, the term *homology*, according to Fitch [19], is the relationship between two characters that have descended, usually with divergence, from a common ancestral character. These characters can be any genic (gene or protein sequence), structural (i.e. morphological) or behavioural feature of an organism.

Cladograms, a particular case of dendograms, are branched diagrams that illustrate patterns of relationships, where the branch lengths are not necessarily proportional to the evolutionary time between groups [17]. A phylogenetic tree can thus be seen as a specific type of cladogram where the branch lengths can represent evolutionary time between groups.

Phylogenetic trees can represent evolutionary relationships by rooted or unrooted binary trees [43]. Figure 4.5 shows the topology of an unrooted and a rooted phylogenetic tree which consists of vertices of degree 1 called *leaves* and unlabeled *internal* vertices of degree 3. Formally, we say that a vertex v_1 is a descendant of another vertex v, if v lies on the path between v_1 and the root vertex. Edges adjacent to a leaf are called *pendant* edges, while all other edges are *internal*. On the other hand, the *edge length* is the distance between nodes

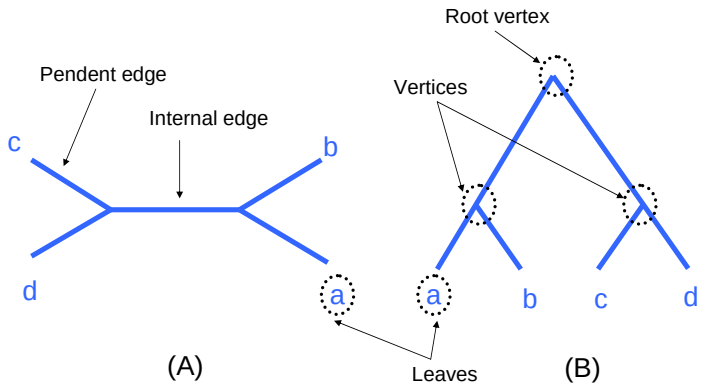


Figure 4.5: Topology and terminology of phylogenetic trees: (A) Unrooted binary tree with four leaves (B) Rooted binary tree with four leaves

In order to create a phylogenetic tree, we need to know the evolutionary distance between the protein sequences. Distance methods aim to construct an all-to-all matrix from the set of sequences describing the distance between each sequence pair. Often, the metric is chosen to be the Euclidean distance. These methods are based on the idea of grouping together those two sequences that are closest first, recalculating the distances, and then grouping again in an iterative process which stops when all the sequences have been grouped. Importantly, the order in which the sequences are clustered determines the graph topology [17].

As a result of the grouping procedure, the phylogenetic tree places the related sequences close together under the same interior node, with the branch lengths closely reproducing the observed distances between sequences.

4.4.2 Interpretability and concordance with the KGTM

In this section, we explore the phylogenetic tree structure of the 232 GPCR amino acid sequences already grouped and visualized using KGTM. With this, we aimed to find whether the data structure revealed by KGTM properly reflects the phylogenetic structure of the data.

In our experiments, phylogenetic trees were obtained using the Java alignment editor application **Jalview 2.6.1** [60]. In this application, sequences are introduced in FASTA format and the trees are calculated on the basis of a measure of similarity between each pair of sequences in the alignment.

The BLOSUM62 scoring matrix [24] was used as the basis for the application of the Unweighted Pair-Group Method with Arithmetic Mean (UPGMA) algorithm [57], which examines the structure present in a similarity matrix and builds the corresponding phylogenetic tree.

The BLOSUM62 (BLOcks of Amino Acid SUbstitution Matrix) is a scoring system [15] to obtain the best sequences alignment based on the 20×20 BLOSUM 62 scoring matrix, which compares the sequences with no less than 62% divergence. During the process, every possible identity and substitution is assigned a score based on the observed frequencies of such occurrences in alignments of related proteins. BLOSUM with high numbers are used for highly related sequences, while low numbers are used for distantly related proteins. Thus, complete identities are assigned the most positive scores. Frequently observed substitutions also receive positive scores and seldom observed substitutions are given negative scores (See example in figure 4.6).

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 4.6: A BLOSUM 62 scoring matrix example

The UPGMA algorithm can be summarized as follows [14]:

1. **Initialisation:**

   - Define n clusters $C_1, ..., C_n$, one for each sequence
   - Initial tree $T = \{leave\ nodes\}$, all at height zero

2. **Iteration:**

   - Define distance between cluster C_k and C_l as the average distance between all pairs of sequences from each cluster:

$$\delta_{kl} = \frac{1}{|C_k|\,|C_l|} \sum_{i \in C_k} \sum_{j \in C_l} d\,(i, j)$$

   - Determine two clusters, k and l for which δ_{kl} is smallest.
   - Define a new cluster $C_m = C_k \cup C_l$ and remove clusters C_k and C_l.
   - Updating tree: add a node m with daughter nodes k and l to the tree T and place it at height $\frac{\delta_{kl}}{2}$.

3. **Termination:**

   - When only two clusters k and l remain place the root node at height $\frac{\delta_{kl}}{2}$.

The pairwise distances which are visualized in the phylogenetic trees show the relationships between the protein sequences in each GPCR group.

Ultimately, UPGMA yields a distance-based sequence clustering solution in the same sense that KGTM provides one. There are radical differences between them, though. UPGMA is strictly hierarchical in nature and proceeds agglomeratively. It means that once agglomerated, clusters cannot be partitioned any longer throughout the procedure. This introduces a directional bias in the solution. Also importantly, cluster assignments at each level of the tree hierarchy are completely symmetrical; that is, the relative position of a sequence within each cluster is arbitrary, which makes the direct interpretation of proximity not too straightforward, specially for big trees.

On the other hand, KGTM (at least as presented in this thesis) is not hierarchical or agglomerative in nature, which avoids any directional bias. Also, its visualization map makes the assessment of proximity far more intuitive and devoid of any symmetry-related artifacts.

In the comparative results reported in the following figures, and due to page size limitations, phylogenetic trees can be only partially represented. The tree parts displayed are those that contained the data associated to each KGTM partial map. For the visualization of the complete phylogenetic tree, see figure B.9 in appendix B and a summarized version of the tree is visualized in the figure B.10.

The visual comparison of the phylogenetic tree with the KGTM mode projection of the data, confirms that the overall grouping structure is quite similar. For example, figure 4.7 reveals that the sequences corresponding to type 1 grouped by the tree in a way that is quite consistent with KGTM. See, for instances how

the KGTM clusters 51 and 61, or 81 and 91 are also close together in the tree. The same can be said about clusters 73, 83, and 93 which are contiguous in the tree and understood by KGTM as a differentiated subgroup. Notice though that the KGTM reveals a mixing of types 1 and 6 in cluster 81 that is by no means obvious in the tree. This is, in fact an undesired effect of the tree branching symmetry, which ocludes a direct assessment of this proximity. Clusters 80 and 98 have a very mixed nature, and only the latter is type 1-dominated. This is clear in the tree, where both clusters can be seen in the neighbourhood of types 6 and 7. Type 8 can be seen in the neighbourhood of the type 1 tree location, but its boundary-like location is far more obvious from the KGTM visualization.

This boundary-like type 8 (See figure 4.8, showing neighbouring relations with types 1, 4 and 6) is pure in its composition, which is an indication that is a more radically different type of GPCR. Interestingly, this is rather obscured in the phylogenetic tree, where the locations associated to type 8 are separated by locations assigned to types 1 and 6. Again, this may just be a byproduct of branching symmetry that can only be assessed through detailed inspection of the tree construction. The KGTM has also revealed some sub-structure within this type, which is at least partially corroborated by the tree solution. This KGTM sub-structure seems to separate the small clusters located at the top of the KGTM map from those at the bottom (which correspond to two neatly separated branches of the tree), with big cluster 56 somewhere in the middle. The exception to this interpretation are the small clusters 36 and 39.

The KGTM visualization reveals that the GPCR sequences most radically separated from the rest are those belonging to type 4 (see figure 4.9), also very pure in its composition. This is neatly reflected in the phylogenetic tree, which locates type 4 in very isolated branches, with type 8 as the only neighbouring relation. Also, the clusters detected by KGTM are extremely coherent with the three representation of the sequences assigned to them.

Figure 4.7: Complete matched visualization of Type 1 including the KGTM mode projections and the corresponding phylogenetic tree.

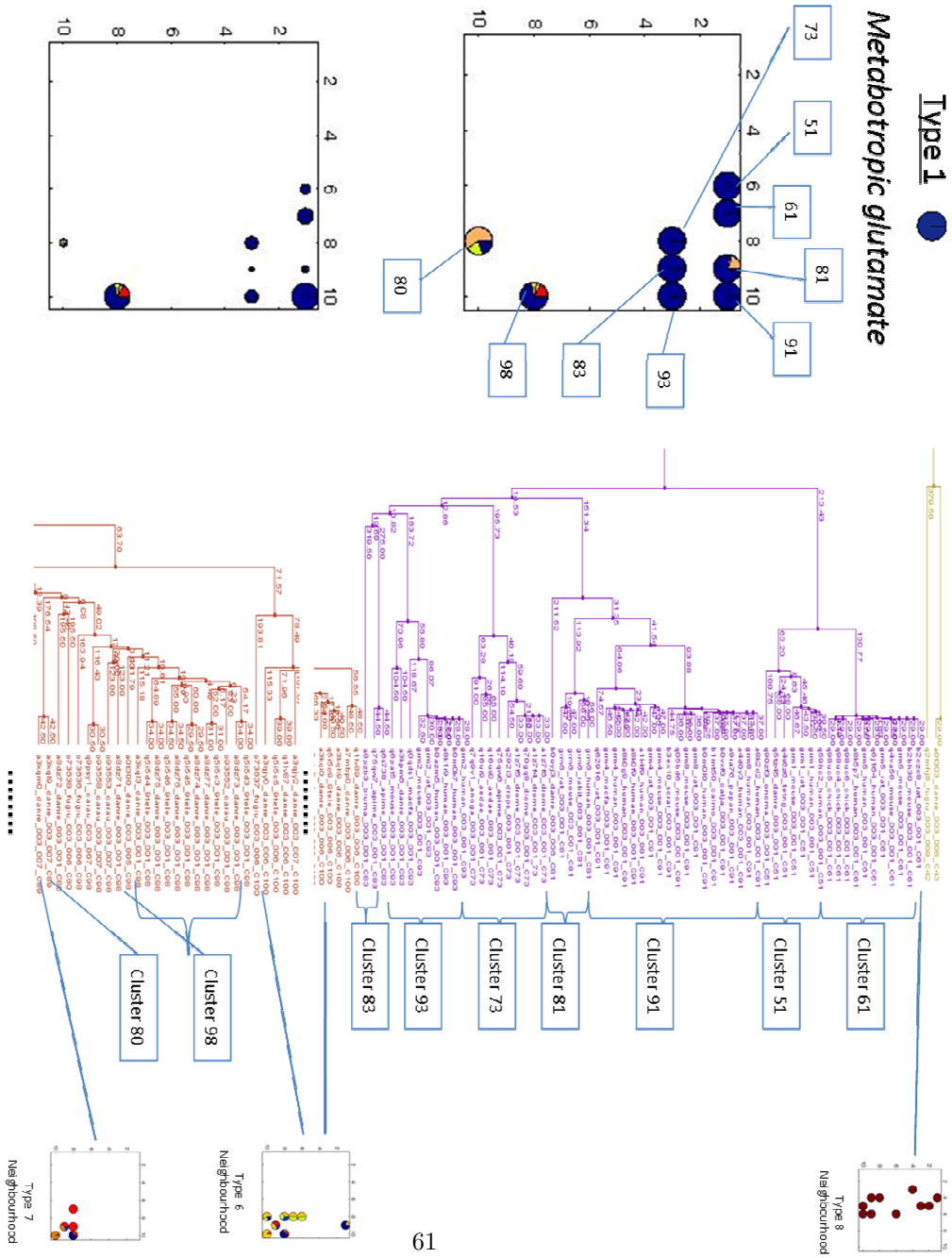


Figure 4.8: Complete matched visualization of Type 8 including the KGTM mode projections and the corresponding phylogenetic tree.

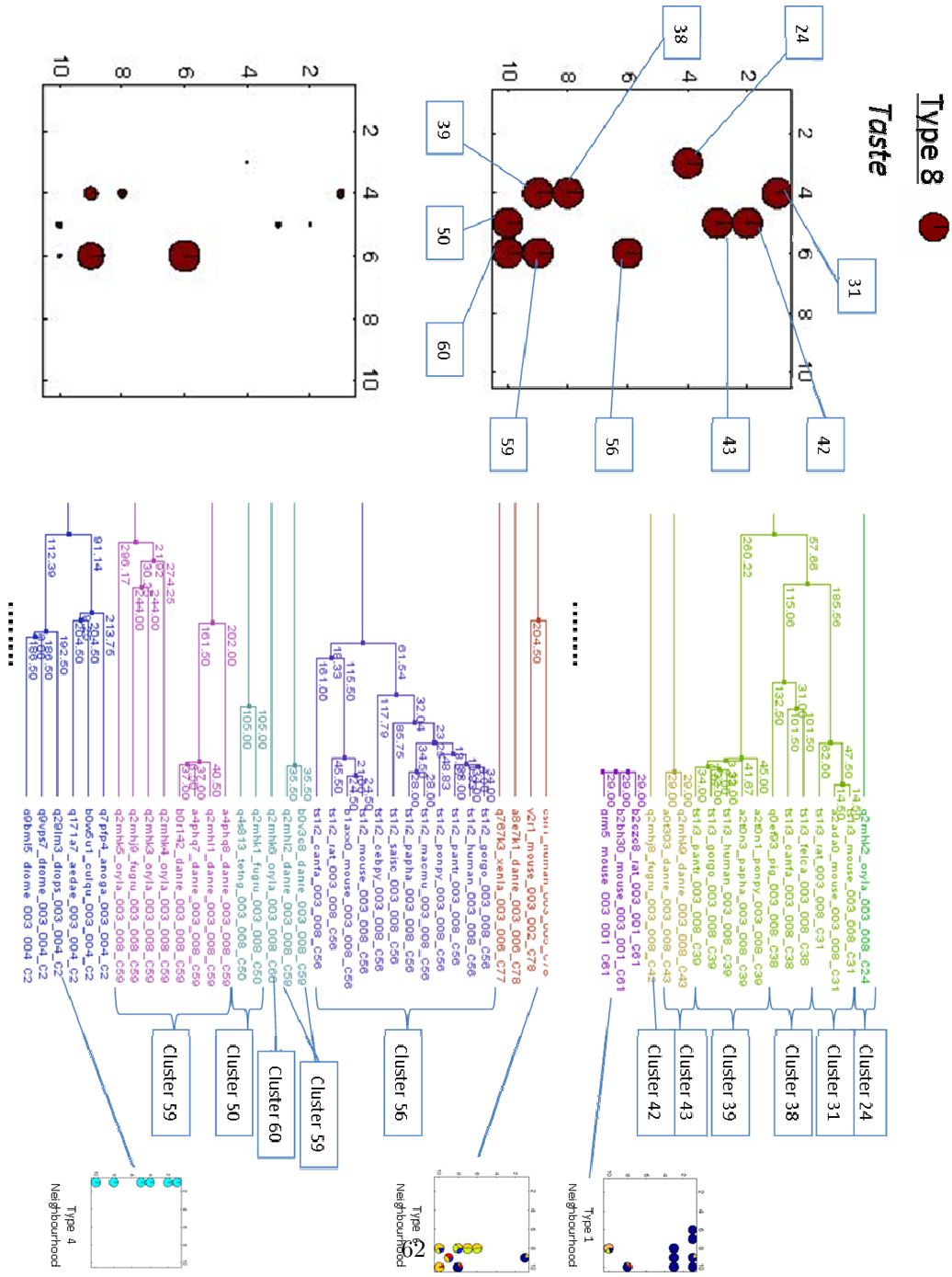
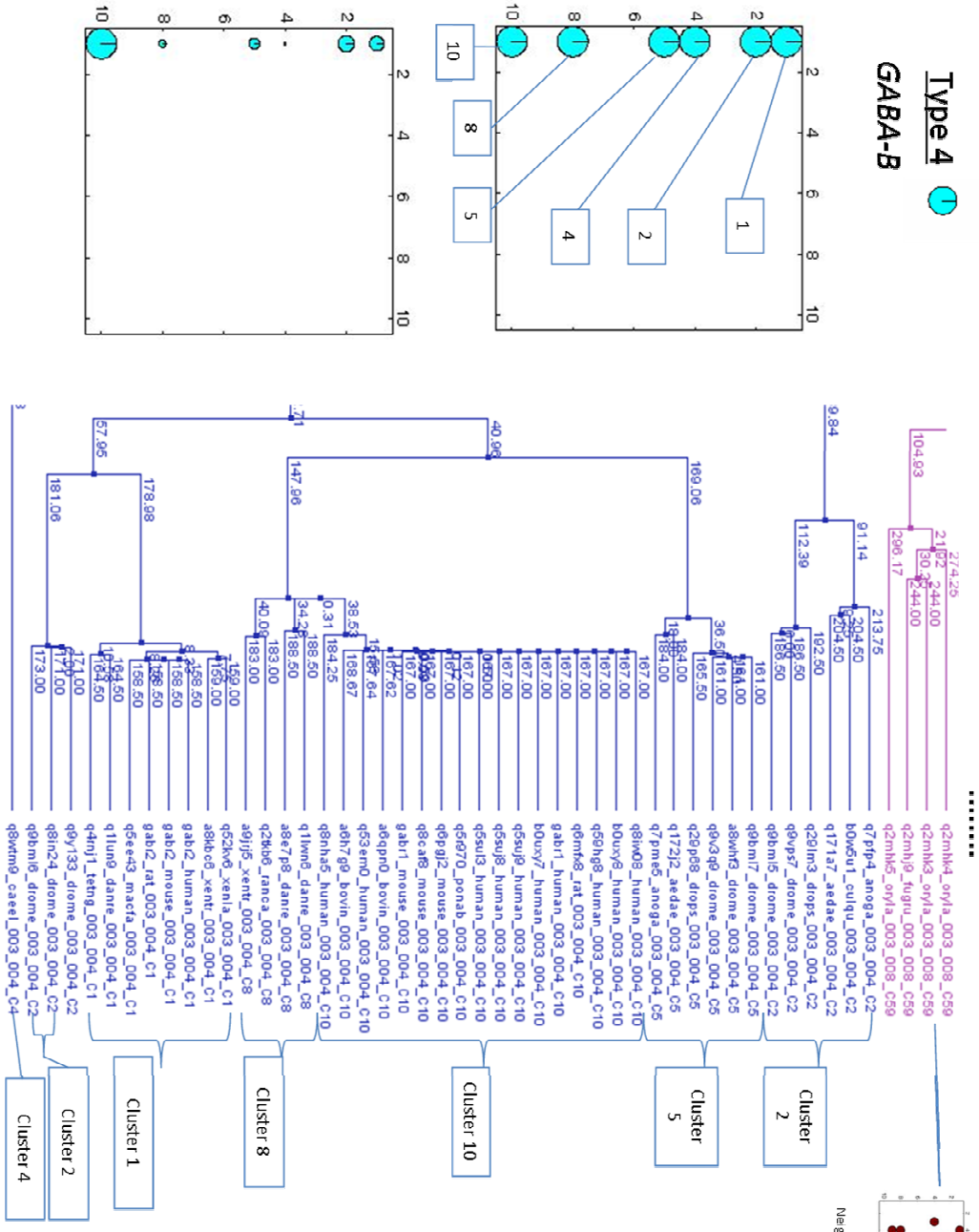


Figure 4.9: Complete matched visualization of Type 4 including the KGTM mode projections and the corresponding phylogenetic tree.



Finally, types 2, 5, 6 and 7 show a far more heterogeneous structure that is also revealed in the tree. They are depicted in figures 4.10 to 4.13, and a more detailed description is omitted here. Overall, the KGTM has been found to provide a neat and simple, while very informative visualization, which is also coherent with the detailed structure provided by the phylogenetic tree. Thus, the KGTM could be recommended as a first-stage exploratory tool to which the phylogenetic trees can provide a second-stage layer of finely detailed information.

Figure 4.10: Complete matched visualization of Type 2 including the KGTM mode projections and the corresponding phylogenetic tree.

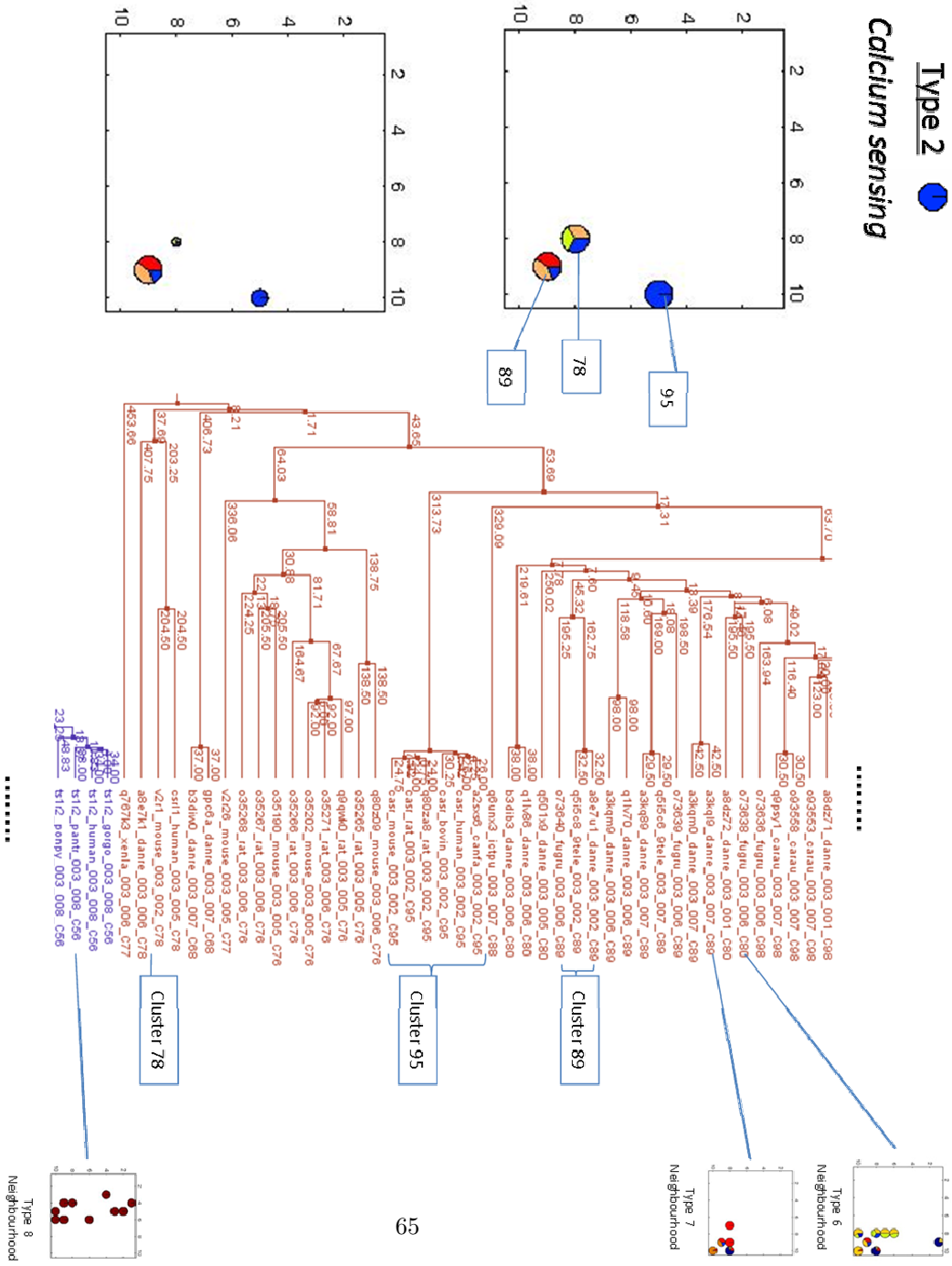


Figure 4.11: Complete matched visualization of Type 5 including the KGTM mode projections and the corresponding phylogenetic tree.

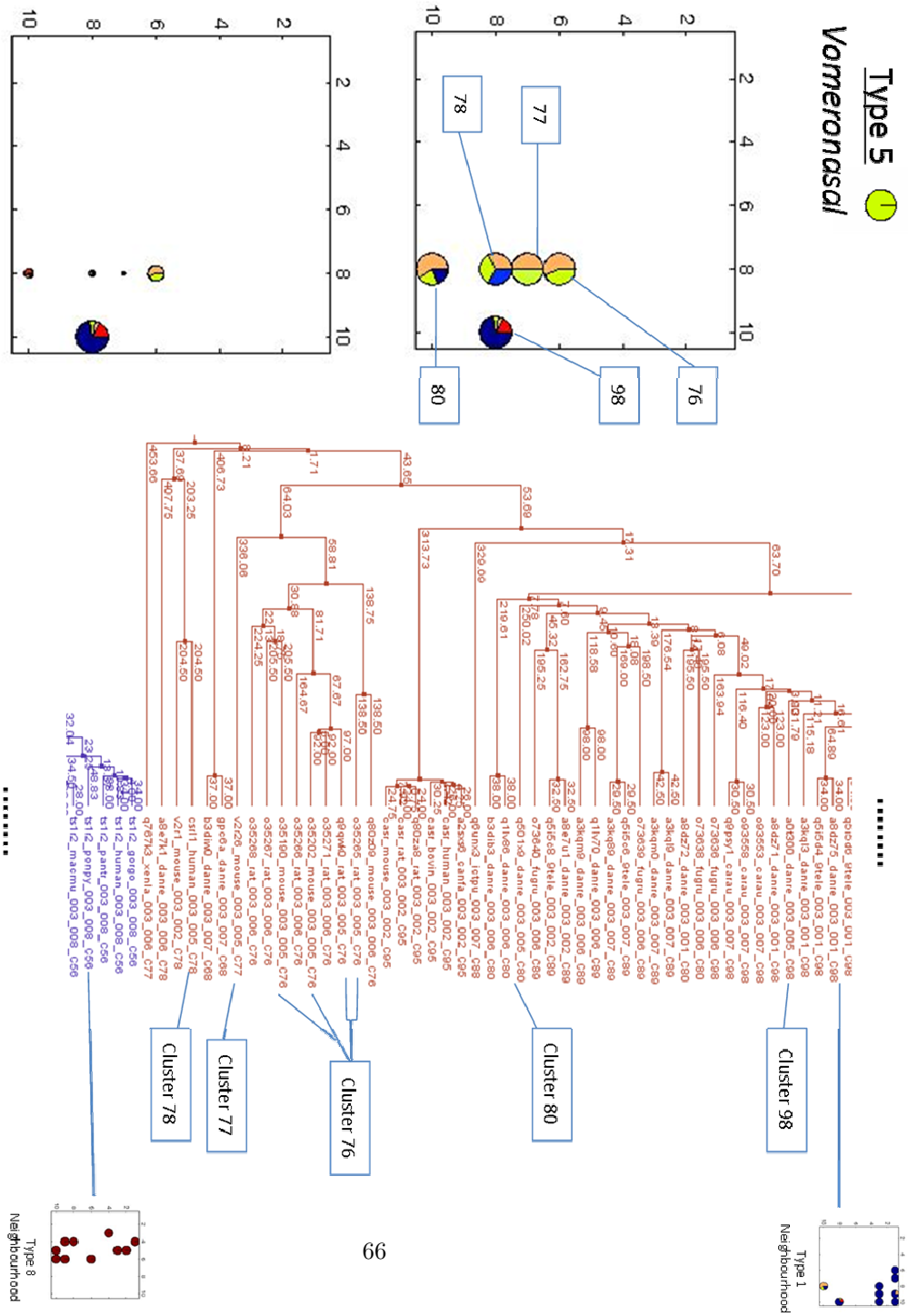


Figure 4.12: Complete matched visualization of Type 6 including the KGTM mode projections and the corresponding phylogenetic tree.

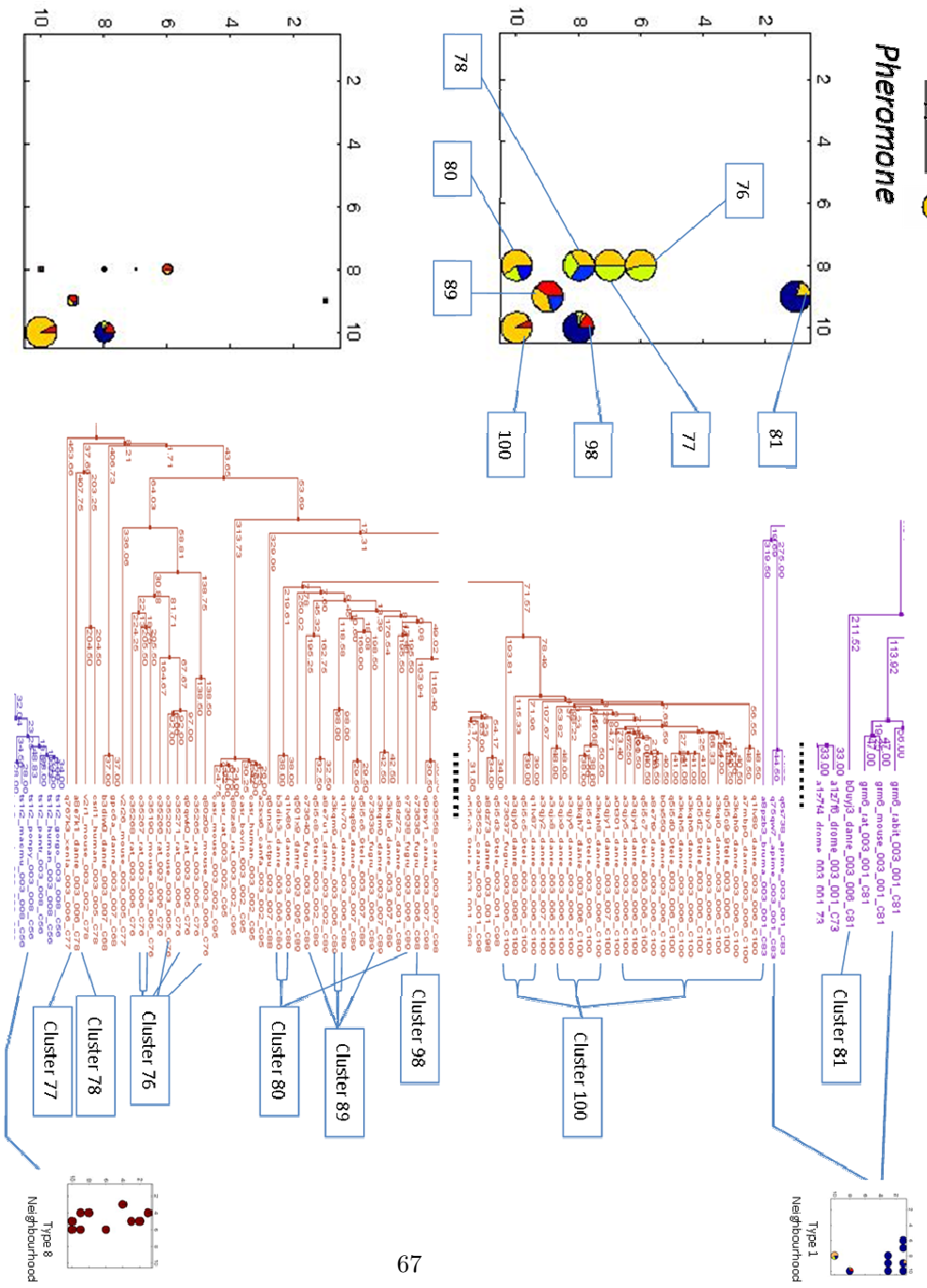
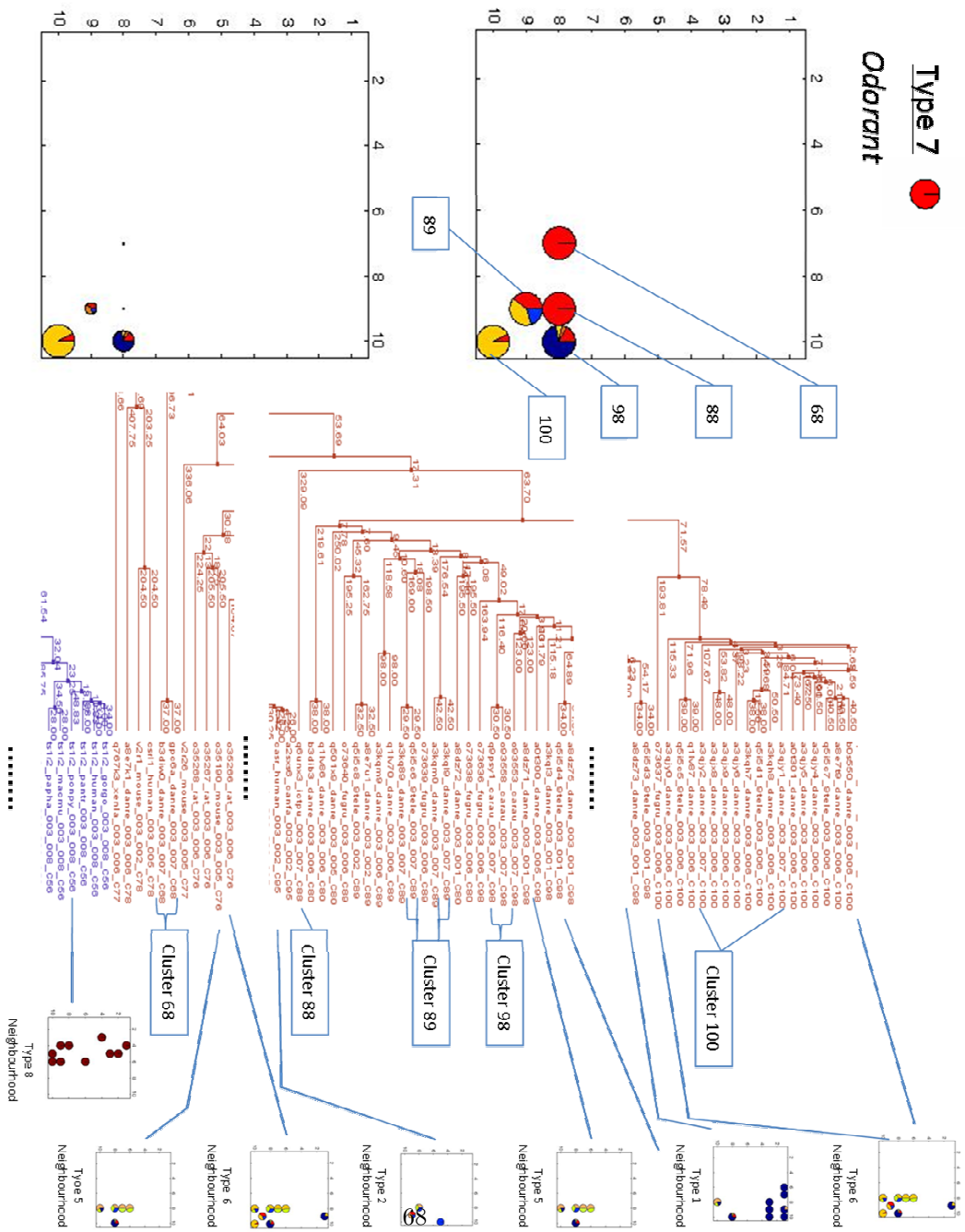


Figure 4.13: Complete matched visualization of Type 7 including the KGTM mode projections and the corresponding phylogenetic tree.



Chapter 5

Conclusions and future work

The world of pharmacology is pointedly veering towards research based on the data generated by pharmacogenomics and proteomics. More than half of the existing drugs target just a handful of protein families and much research in the area is currently devoted to analysing protein amino acid sequences.

The unravelling of the three-dimensional structure of GPCR's would be key for the understanding of its function and, therefore, for their applicability in pharmacological research. This is a hard and mostly unresolved problem. On the meantime, their amino acid sequences are easy to acquire and widely available. The grouping of GPCRs into types and subtypes based on the analysis of these sequences is a building block towards their full characterization.

In this brief study, we have shown a kernel method of the manifold learning family, namely the KGTM, which is capable of simultaneously revealing the grouping structure of GPCRs while making the intuitive visualization of such structure possible. This grouping problem is most commonly dealt with in the field using phylogenetic trees. Phylogenetic trees are a widely used graphical

tool in the field of proteomics, and they visually illustrate through hierarchical dendograms the evolutionary closeness between sequences.

The results reported in the previous chapter and complemented by the appendices indicate that the KGTM yields a quite clear (and intuitively interpretable) grouping structure for the GPCR Family C types. Several subtypes occupy quite differentiated areas on the map, showing little overlapping. A few of them, instead, have at least partially overlapping representations. We have focused on one of these types, the metabotropic glutamate receptors, due to the fact that they play important roles in regulating cell excitability and synaptic transmission, being widely distributed throughout the central nervous system. This makes them an important pharmacological target for a whole range of neurological and psychiatric disorders. The KGTM has been shown to provide a neat representation of the fine structure of these subtypes.

Importantly, the distribution of the groups revealed by the phylogenetic tree of the analysed data shows striking coincidences with the groupings yielded by KGTM, even, sometimes, down to the fine detail. This is a fine example of how kernel methods for learning in structured domains can be useful in a biological application context. The phylogenetic tree introduces a directional bias in the grouping solution. Moreover, cluster assignments at each level of the tree hierarchy are completely symmetrical, which makes the direct interpretation of proximity not difficult to achieve. The KGTM visualization map makes the assessment of proximity far more intuitive and devoid of any symmetry-related artifacts.

The KGTM, as applied in this thesis, lacks hierarchical structure (although hierarchies can be inferred and, to some extent, have been illustrated with some of the analysed data). Thus, the KGTM, even if not a tool to replace phylogenetic trees, could confidently be recommended as a first-stage exploratory tool to which the phylogenetic trees can provide a second-stage layer of finely detailed information. In other words, both tools could complement each other,

amplifying their individual advantages.

A straightforward extension of KGTM for future research is precisely its definition within a hierarchical framework. For this, we could find inspiration in previous work concerning similar models such as SOM [21], mixture models [5], [61], [59], or even the own GTM [58]. The obtained hierarchical structure could be compared to that provided by phylogenetic trees in a more principled way.

In future research, KGTM might be used to help in the exploration of receptors with very heterogeneous grouping structure. This heterogeneity might be a clue to their susceptibility towards heterodimerization, which could be useful in the quest of more potent and safer drugs. The KGTM could also offer the possibility of detecting receptors that are either misclassified (and thus be used for database curation) or not assigned in the original database, a situation commonly known as "receptor deorphanization".

Chapter 6

Thesis publications

1. Vellido A., Cárdenas M.I., Olier I., Rovira X. and Giraldo J. (2011) A probabilistic approach to the visual exploration of G Protein-Coupled Receptor sequences. In Procs. of the 19th European Symposium on Artificial Neural Networks (ESANN 2011), 233-238.

2. Cárdenas M.I., Vellido A., Olier I., Rovira X. and Giraldo J. (2011)A kernel-based visualization of g protein-coupled receptor sequences closely resembles their phylogenetic tree. Eight International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2011), Gargnano-Lago di Garda, Italy *(To be published as part of a volume of Lecture Notes in Bioinformatics (LNBI), Springer).*

3. Cárdenas M.I., Vellido A., Olier I., Rovira X. and Giraldo J. (2011) Analyzing Protein Sequences for Pharmacology. In: Intelligent Data Analysis for Real-Life Applications: Theory and Practice, Magdalena R., Martínez M., Martínez J.M., Escandell P. and Vila J.(Eds.), IGI Global. *Forthcoming.*

Bibliography

[1] Andras P. (2002). Kernel-Kohonen networks. *International Journal of Neural Systems*, Volumen 12, p.117-135.

[2] Baldi P., Brunak S. (2001). Bioinformatics: The Machine Learning Approach. MIT Press.

[3] Barnes P.J. (2006) Receptor heterodimerization a new level of cross-talk. *Journal of Clinical Investigation*, 116(5), 1210-1212.

[4] Bishop C.M., Svensén M. and Williams C. K. I. (1998) GTM: The Generative Topographic Mapping, *Neural Computation, Elsevier*, 10(1),215-234.

[5] Bishop C.M. and Tipping M.E. (1998) A Hierarchical Latent Variable Model for Data Visualisation, IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 3(20),281-293.

[6] Bishop C. M., Svensén M. and Williams C. K. I. (1998) Developments of the Generative Topographic Mapping, *Neurocomputing*, 21(1-3),203-224.

[7] Choi J.Y., Qiu J., Pierce M. and Fox G.(2010) GTM by deterministic annealing. *International Conference on Computer Science ICCS*, 1(1), 47-56.

[8] Clamp M., Cuff J., Searle S.M. and Barton G.J. (2004) The Jalview Java alignment editor. *Bioinformatics Applications Note.* 20(3), 426-427.

[9] Cobanoglu M. C., Saygin Y. and Sezerman U. (2010) Classification of GPCRs using family specific motifs. *IEEE-ACM Transactions on Computational Biology and Bioinformatics.*, In press.

[10] Conn P.J., Christopoulos A. and Lindsley C.W.(2009) Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nature Reviews. Drug Discovery.* Volume 8.

[11] Dempster A. P., Laird N. M. and Rubin D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, B, 39(1),1-38.

[12] Drews J. (2000)Drug Discovery: A Historical Perspective. *Science* 17. 287 (5460), 1960-1964.

[13] Doumazane E., Scholler P., Zwier J. M., Trinquet E., Rondard P. and Pin J-Ph (2010) A new approach to analyze cell surface protein complexes reveals specific heterodimeric metabotropic glutamate receptors *The FASEB Journal* September 27(10), 66-77.

[14] Durbin R., Eddy S. R., Krogh A., and Mitchison G. (2004) *Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University in Press*, Cambridge.

[15] Eddy S.R. (2004) Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology.* 22(8), 1035-1036.

[16] Ellis C. (2004) The state of GPCR research in 2004, *Nature Reviews Drug Discovery* 3, 577-626. July.

[17] Felsenstein J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology.*, 266:418-27.

[18] Filmore D. (2004) Cell-based screening assays and structural studies are fueling G-protein coupled receptors as one of the most popular classes of investigational drug targets. Volume 7 (11), *Modern Drug Discovery*, November.

[19] Fitch W.M. (2000) Homology: a personal view on some of the problems. *Elsevier Science.* TIG. Volume 5. May.227-231.

[20] Foreman J. C. and Johansen T. (2003) Textbook of receptor pharmacology. CRC Press, 2nd edition.

[21] Furukawa T.(2009) SOM of SOMs. *Neural Networks* 22 (4). May.463-478

[22] Gilman A.G. (1987) G proteins: transducers of receptor-generated signals. *Annual Review of Biochemistry.* 56,615-649.

[23] Goudet C., Binet V. and Prezeau L. and Pin J-Ph. (2004) Allosteric modulators of class-C G-protein-coupled receptors open new possibilities for therapeutic application *Drug Discovery Today: Therapeutic Strategies* Vol. 1, No. 1.

[24] Henikoff S. (1992) Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences PNAS* 89, 10915-10919.

[25] Heskes T. (1998). Energy functions for self-organizing maps. Energy functions for self-organizing maps, *Theoretical Foundation SNN*, University of Nijmegen.

[26] Horn F., Weare J., Beukers M. W., Horsch S., Bairoch A., Chen W., Edvardsen O., Campagne F. and Vriend G. (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Research* 26, 275-279.

[27] Jolliffe I.T. (2002) Principal Component Analysis, *Springer Series in Statistics (2nd edition)*. Springer, NY.

[28] Julio-Pieper M., Flor P. J., Dinan T. G. and Cryan J. F. (2011) Exciting Times beyond the Brain: Metabotropic Glutamate Receptors in Peripheral and Non-Neural Tissues. *Pharmacological Reviews* March 2011, 63(1), 35-58.

[29] Kahn S.D. (2011). On the future of genomic data. *Science*, 331(6018): 728-729.

[30] Kohonen T. (1982) Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59-69.

[31] Kohonen T. (2001) *Self-Organizing Maps (3rd ed)*. Springer-Verlag, Berlin.

[32] Lampinen J. and Oja E. (1992). Clustering properties of hierarchical selforganizing maps. *Journal of Mathematical Imaging and Vision*, 2, 261-272.

[33] Lau K.W., Yin H. and Hubbard S. (2006) Kernel self-organising maps for classification. *Neurocomputing* 69. 2033-2040

[34] Lipman D.J.and Pearson W.R. (1985). Rapid and sensitive protein similarity searches. *Science* 22, March.227 (4693), 1435-1441

[35] Lisboa, P.J.G. (2002). A Review of Evidence of Health Benefit from Artificial Neural Networks in Medical Intervention. *Neural Networks* 15, 9-37

[36] Lisboa P.J.G., Vellido A., Tagliaferri R., Napolitano F., Ceccarelli M., Martin-Guerrero J.D. and Biganzoli E. (2004) Data Mining in Cancer Research, *IEEE Computational Intelligence Magazine*, 5(1), 14-18.

[37] MacDonald D. and Fyfe C.(2000)The Kernel Self Organising Map. *Applied Computational Intelligence Research Unit*, The University of Paisley, Scotland.

[38] MacKay D. J. C. (1992) A Practical Bayesian Framework for Back-propagation Networks. *Neural Computation*, 4(3),448-472.

[39] Mardis E.R. (2011) A decade's perspective on DNA sequencing technology. *Nature* 470, 198-203. February.

[40] Olier I. and Vellido A. (2008) Advances in clustering and visualization of time series using GTM Through Time. *Neural Networks*, *Elsevier*, 21(7), 904-913.

[41] Olier I., Vellido A. and Giraldo J. (2010) Kernel Generative Topographic Mapping. *In Procs. of the 18th European Symposium on Artificial Neural Networks* (ESANN 2010), 481-486.

[42] Overington J. P., Al-Lazikani B. and Hopkins A. L. (2006) How many drug targets are there?, *Nature Reviews, Drug Discovery*, VOLUME 5, December.

[43] Page R.D.M. and E.C. Holmes. (1998) Molecular Evolution: A Phylogenetic Approach. Blackwell Science. ISBN 0865428891. Oxford. Chapter 2. p.11.

[44] Pan Z. S., Chen S. C. and Zhang D. Q. (2004). A kernel-base SOM classifier in input space. *Acta Electronica Sinica*, 32, 227-231 (in Chinese).

[45] Pearson K. (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* 2 (6), 559-572

[46] Pierce K.L., Premont R.T. and Lefkowitz R.J. (2002) Seven-transmembrane receptors. Nature. Reviews. *Molecular Cell Biology*. 3:639-650.

[47] Pin J.P., Galvez T. and Prézeau L. (2003) Evolution, structure and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacology Therapeutics*. June 98(3), 325-354.

[48] Rabiner L.R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. PROCEEDINGS OF THE IEEE. 8825949. VOL. 77, NO. 2, FEBRUARY

[49] Rondard Ph., Liu J., Huang S., Malhaire F., Vol C., Pinault A., Labesse G. and Pin J.P.(2006) Coupling of Agonist Binding to Effector Domain Activation in Metabotropic Glutamate-like Receptors. *Journal of Biological Chemistry*. AUGUST 25. 281 (34).

[50] Rondard Ph., Goudet C., Kniazeff J., Pin J-Ph. and Prézeau L. (2011) The complexity of their activation mechanism opens new possibilities for the modulation of mGlu and GABAB class C G protein-coupled receptors. *Neuropharmacology* 60. 82-92

[51] Rovira X., Pin J.P. and Giraldo J. (2010) The asymmetric/symmetric activation of GPCR dimers as a possible mechanistic rationale for multiple signalling pathways. *Trends Pharmacology Science*. January. 31(1),15-21.

[52] Rovira X. (2010) PhD Thesis: Assessing the functionality of G protein-coupled receptor oligomerization with mathematical modeling. Institut de Neurociències, Unitat de Bioestadística, Universitat Autònoma de Barcelona, 08193, Bellaterra (Barcelona) - Spain.

[53] Schölkopf B., Smola A. and Müller K.R. (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10, 1299-1319.

[54] Schölkopf B. and Smola A. (2002) *Learning with Kernels*. The MIT Press, Cambridge, Massachusetts.

[55] Schölkopf B., Tsuda K. and Vert J-Ph. (2004) Kernel Methods in Computational Biology, MIT Press.

[56] Shawe-Taylor J. and Cristianini N. (2004) Kernel Methods for Pattern Analysis. Cambridge University Press.

[57] Sokal R. and Michener C. (1958) A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 38: 1409-1438.

[58] Tino P. and Nabney I. (2002) Hierarchical GTM: Constructing Localized Nonlinear Projection Manifolds in a Principled Way, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5) p.639-656

[59] Vasconcelos N. and Lippmann A.(1999) Learning Mixture Hierarchies. In: Advances in NIPS 11, p.606-612

[60] Waterhouse A.M., Procter J.B., Martin D.M.A, Clamp M. and Barton G. J. (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25 (9) 1189-1191

[61] Williams C. (2000) A MCMC Approach to Hierarchical Mixture Modelling. In: Advances in NIPS 12, p. 680-686

[62] Yin H. (2008) The Self-Organizing Maps: Background, Theories, Extensions and Applications, *Studies in Computational Intelligence* (SCI) 115, 715-762.

[63] Yang Z.R. and Thomson R. (2005) A novel neural network method in mining molecular sequence data, *IEEE Transactions on Neural Networks.*, 16:263-274.

Appendix A

G-Protein coupled receptors included in the data set

Table A.1: Dataset observations obtained from GPCRDB [26].

G-Protein Coupled Receptors ID of the dataset				
$ts1r3\_mouse$	$q7pfp4\_anoga$	$q9bml7\_drome$	$grm1\_caeel$	$b2czc8\_rat$
$a8whf3\_drome$	$a1z7f6\_drome$	$q52kv6\_xenla$	$grm4\_mouse$	$q2mhk9\_danre$
$a1z7f4\_drome$	$q1lv89\_danre$	$grm2\_human$	$q4rxb9\_tetng$	$a3kqh9\_danre$
$q8iw08\_human$	$q5i5d3\_9tele$	$a8dz74\_danre$	$ts1r2\_ponpy$	$a1imb8\_caeel$
$q80z09\_mouse$	$a1imb9\_caeel$	$q62916\_rat$	$a8e7p8\_danre$	$q9qwk0\_rat$
$ts1r3\_felca$	$q5i5d2\_9tele$	$q59hc2\_human$	$ts1r2\_pantr$	$a4phq8\_danre$
$q16us0\_aedae$	$a3kqi0\_danre$	$grm6\_rabit$	$grm2\_rat$	$q75qw6\_apime$
$b0vxf0\_calja$	$b2bh30\_mouse$	$a3kqh8\_danre$	$q0ef93\_pig$	$grm5\_mouse$
$gpc6a\_human$	$a2sxs6\_canfa$	$q2mhk4\_oryla$	$a2t0n1\_ponpy$	$a8dz75\_danre$
Continued on next page				

Table A.1 – continued from previous page				
o73637_fugru	q6s738_apime	b0uxy8_human	q2mhk3_oryla	ts1r2_mouse
q4sr46_tetng	o73638_fugru	q5i5d5_9tele	o73639_fugru	q172j2_aedae
a3qjy2_danre	a2t0n3_papha	gabr2_human	a8k0f9_human	a8wra6_caebr
q4s6z6_tetng	casr_bovin	q59hg8_human	q8wtm9_caeel	a0t301_danre
q93564_caeel	q80za8_rat	a3qjy6_danre	q2mhk1_fugru	a8e7u1_danre
a8dz73_danre	q501x9_danre	a3kql3_danre	a8dz76_danre	q9y133_drome
q5i5d6_9tele	gpc6a_danre	gpc6a_carau	b1mt50_calmo	q70gq8_drome
grm1_human	q2mhk2_oryla	o35271_rat	a1l1t5_danre	q9bml6_drome
q75qw7_apime	q1lwn6_danre	q4spr7_tetng	a3qjx9_danre	o35265_rat
grm1_mouse	gabr2_mouse	ts1r2_human	o93552_carau	b1axx0_mouse
a8wrn3_caebr	grm_drome	q8nha5_human	q6pgj2_mouse	q7q9v1_anoga
a8dz72_danre	q9vps7_drome	q2tkb6_ranca	ts1r3_human	o35267_rat
grm5_human	q2mhl1_danre	q5suj9_human	v2r1_mouse	ts1r2_cebpy
a8e7t9_danre	grm4_macfa	ts1r1_mouse	casr_mouse	q767k3_xenla
q2mhj9_fugru	a8e7k1_danre	q5i5d1_9tele	q1lun9_danre	o35266_rat
b3diw0_danre	grm5_rat	a8k5p7_human	grm8_mouse	q2mhk6_oryla
a0t303_danre	q5suj8_human	b0w5u1_culqu	a7swz2_nemve	q9bml5_drome
grm2_mouse	o93558_carau	q4t849_tetng	casr_rat	q1lv70_danre
a8wsb1_caebr	q5i5d4_9tele	o35268_rat	gpc6a_rat	q5tz45_danre
q8caf8_mouse	o35269_rat	grm6_mouse	a3kqh5_danre	a3qjy0_danre
a3fpk2_caeel	a1z7f5_drome	b0m0k7_human	ts1r3_gorgo	q8in24_drome
q292i3_drops	b3ex10_sorar	q5sul3_human	b0uyj3_danre	q05bd6_mouse
q5r970_ponab	gabr2_rat	o73640_fugru	q5i5c3_9tele	a6qpn0_bovin
a6h7g9_bovin	ts1r2_canfa	casr_human	o93553_carau	v2r26_mouse
a9jrj5_xentr	a8k1r9_human	a3kqm0_danre	q20c73_drovi	q1lv87_danre
b0r142_danre	a3qjy1_danre	q98uc5_chick	ts1r2_rat	q5i5d0_9tele
Continued on next page				

Table A.1 – continued from previous page				
b0azm5_human	q5i5c6_9tele	a3qjx8_danre	o35202_mouse	q171a7_aedae
a7mbp0_danre	ts1r3_pantr	a3kq89_danre	a4d0y3_human	a0t300_danre
q4spr4_tetng	b2rmx0_human	q5i5c8_9tele	q2mhk5_oryla	q9psy1_carau
q2mhl2_danre	b2re49_human	a8k2d2_human	ts1r2_gorgo	b0v3c8_danre
q29lm3_drops	a3qjy3_danre	ts1r1_rat	q4s834_tetng	a1l4f9_human
grm6_human	q29p68_drops	a2ada0_mouse	a9ra76_papan	q90zf3_oncma
gpc6a_mouse	a2t0n2_saisc	b0s550_danre	q6j164_human	o73636_fugru
a3kqh6_danre	q2mhk0_fugru	ts1r3_canfa	a3kpm6_danre	q6unx3_ictpu
b0m0l0_human	ts1r2_macmu	q4va56_mouse	grm1_rat	q6mfx8_rat
gabr1_human	a8dz71_danre	b0uxy7_human	q4s813_tetng	a8pzb3_bruma
grm4_rat	a8k0g9_human	ts1r3_rat	csrl1_human	o35190_mouse
grm8_rat	q4s836_tetng	q16ur9_aedae	q4rjz9_tetng	a4phq7_danre
grm8_human	grm4_human	q9v3q9_drome	ts1r1_human	q0qds1_macfa
q4rnj1_tetng	a3qjy4_danre	q98uc6_chick	a3kql9_danre	q8mxu2_caeel
q2mhj8_fugru	a3kqh7_danre	q20073_caeel	a3kqm9_danre	ts1r2_saisc
a3qjy5_danre	ts1r2_papha	q5i5c7_9tele	q3u5h1_mouse	a8dz77_danre
q4sfl1_tetng	q5i5c9_9tele	q98uc4_chick	q5ee43_macfa	q5i5c5_9tele
gabr1_mouse	q7pme5_anoga	q1lv86_danre	b3dib3_danre	q53em0_human
grm6_rat	a8kbc6_xentr	q6pat5_mouse		

Appendix B

KGTM visualization of GPCR Family C types

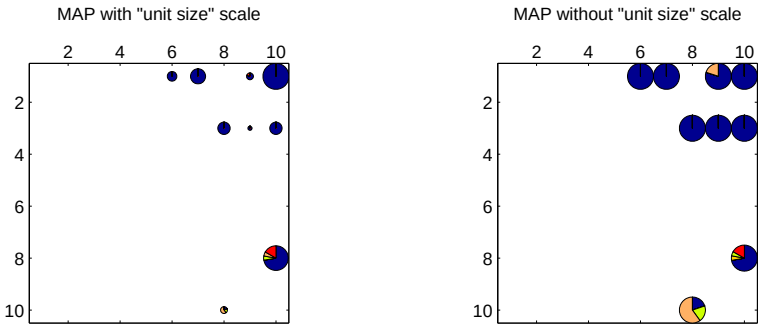


Figure B.1: Type 1 data visualization on a 10×10 KGTM representation map, using the mode projection.

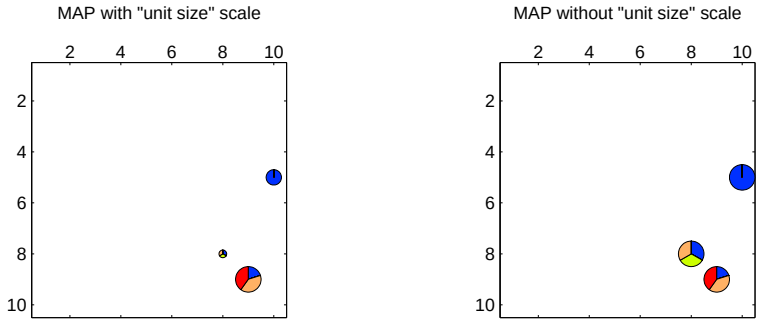


Figure B.2: Type 2 data visualization on a 10×10 KGTM representation map, using the mode projection.

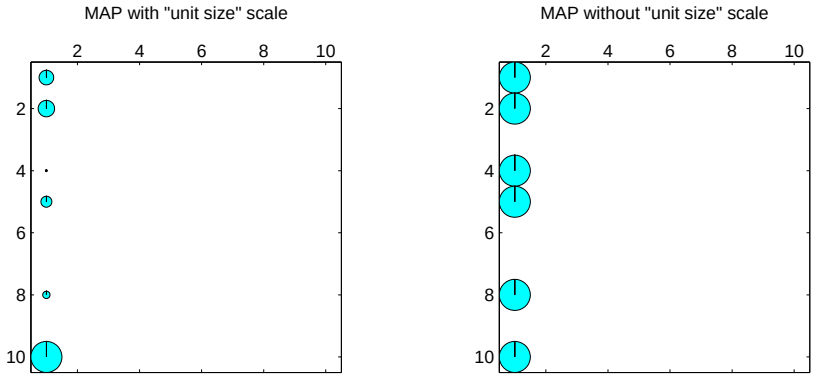


Figure B.3: Type 4 data visualization on a 10×10 KGTM representation map, using the mode projection.

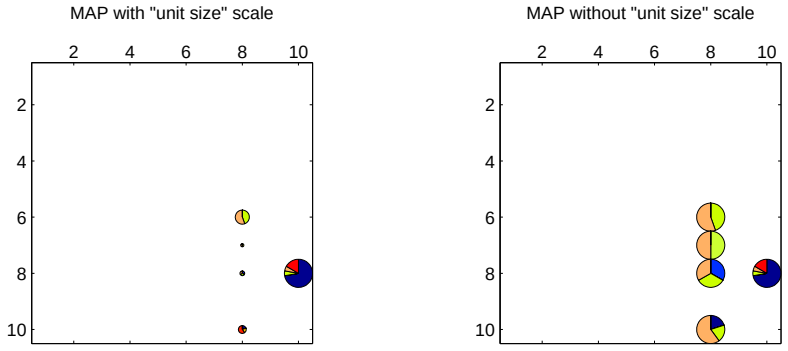


Figure B.4: Type 5 data visualization on a 10×10 KGTM representation map, using the mode projection.

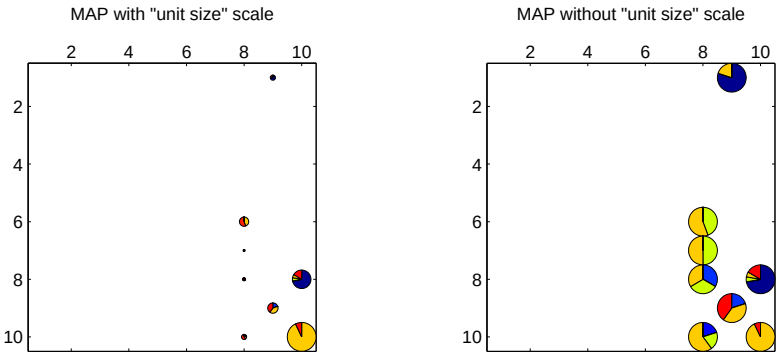


Figure B.5: Type 6 data visualization on a 10×10 KGTM representation map, using the mode projection.

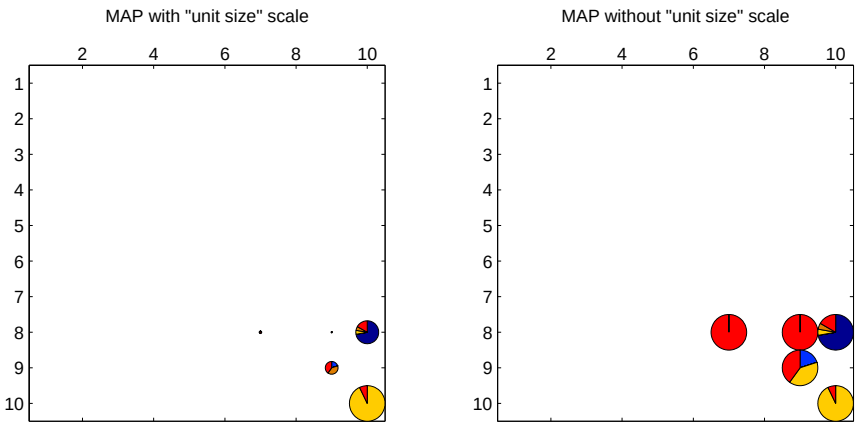


Figure B.6: Type 7 data visualization on a 10×10 KGTM representation map, using the mode projection.

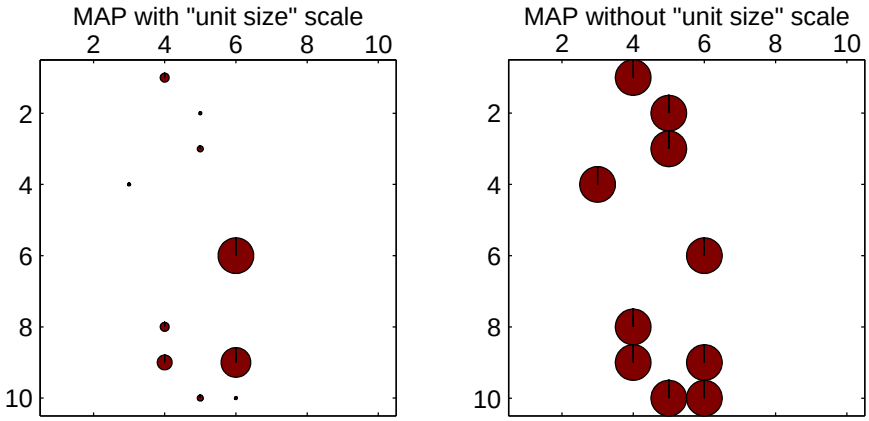


Figure B.7: Type 8 data visualization on a 10×10 KGTM representation map, using the mode projection.

Figure B.8: General hierarchical visualization of GPCR Family C types, including detailed subtyping of mGlu receptors.

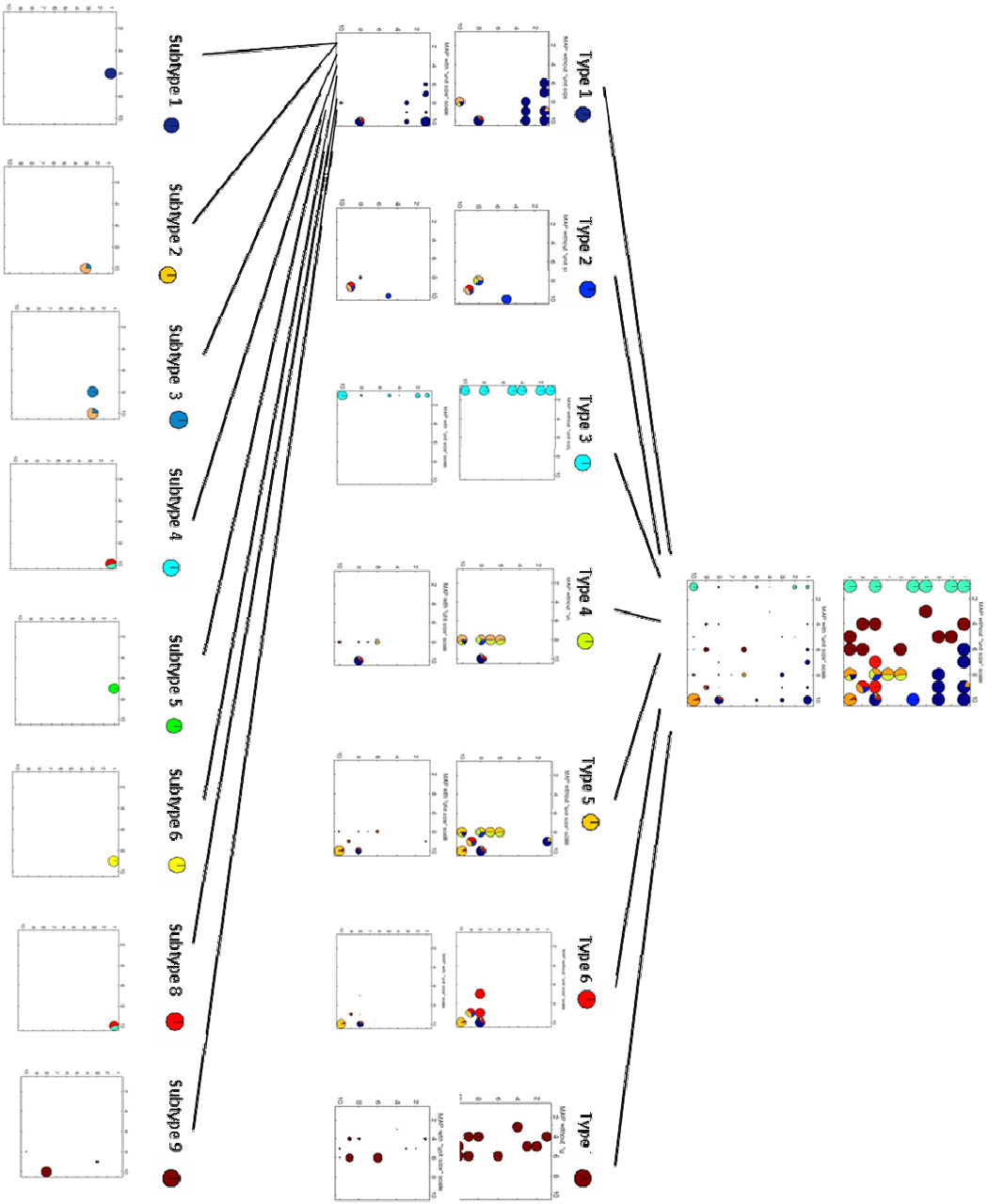
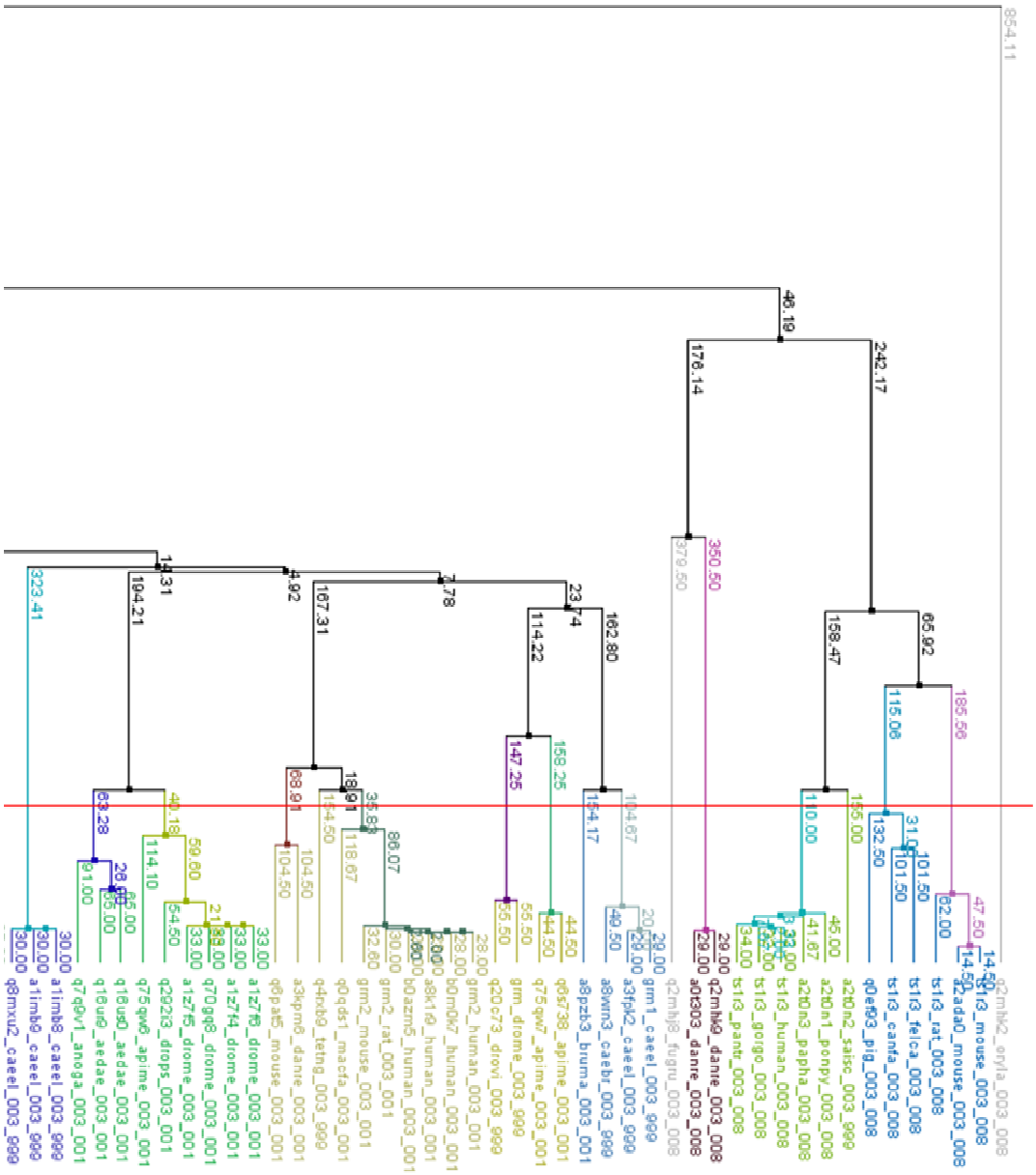
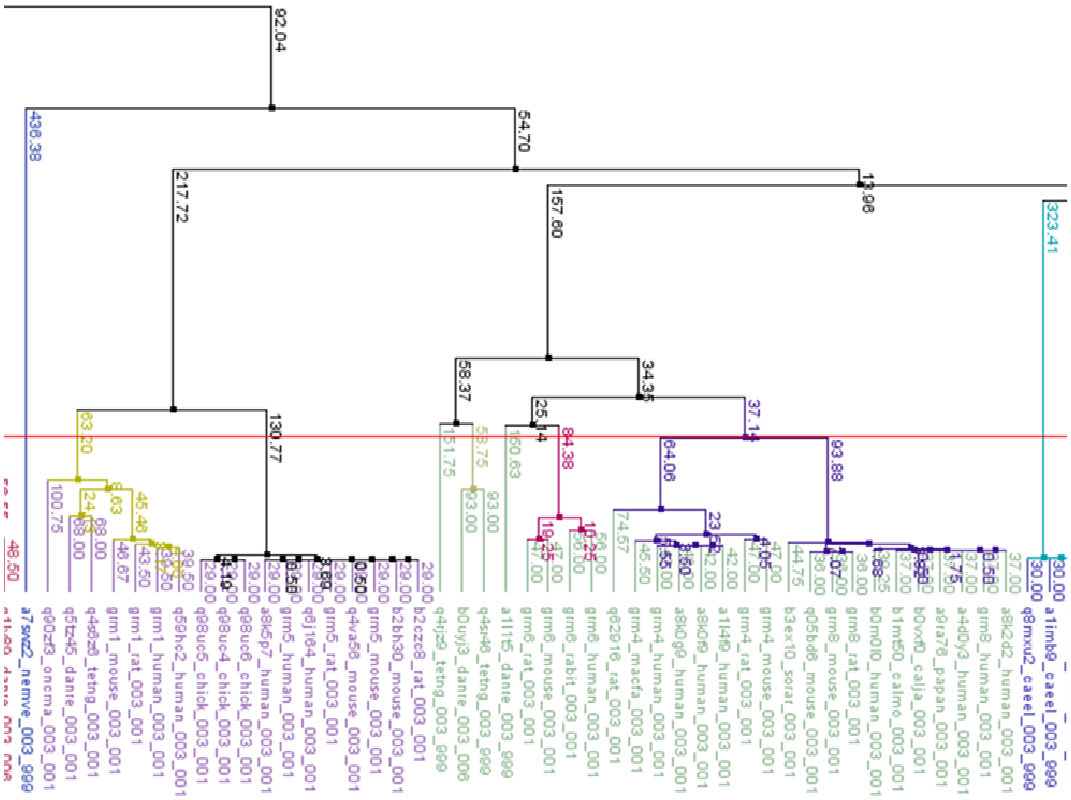


Figure B.9 shows the complete phylogenetic tree representation of the GPCR family C data analysed in this thesis. The colors in the tree are automatically generated by the software. Same color is assigned to close leaves (sequences) and branches (groups of sequences) of the tree, according to the evolutive distance between sequences. This distances are the numbers attached to the branches. The software also automatically plots a red line which establishes the depth from which the color grouping starts. Individual sequences in the leaves of the tree are labelled according to three items: their ID , the family and the type (e.g., sequence $ts1r3\_mouse\_003\_001$ indicates ID: $ts1r3\_mouse$; family: 003 (C); and type: 001).

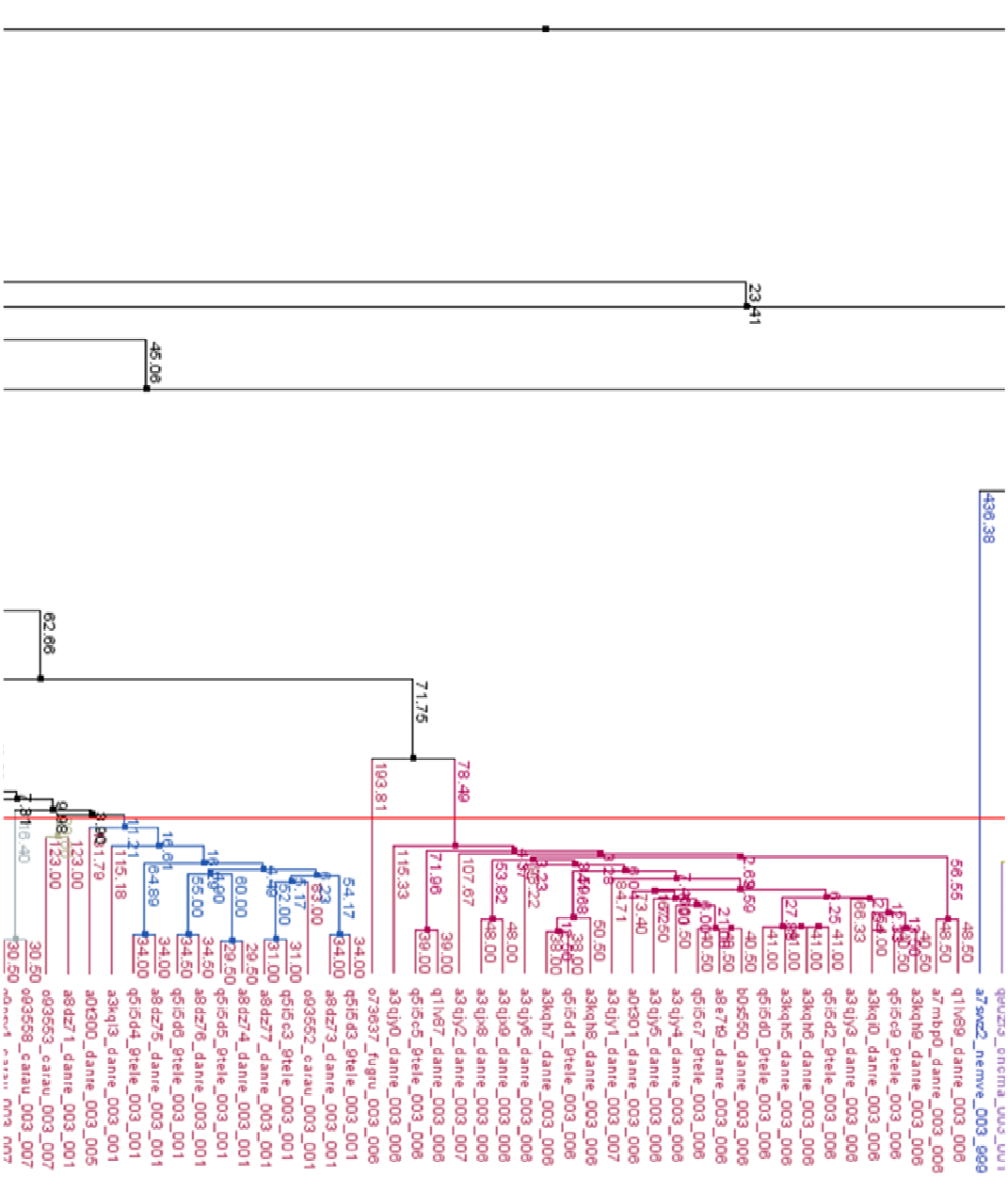
Figure B.9: The figure has been split due to space limitations. - continues on the next page -



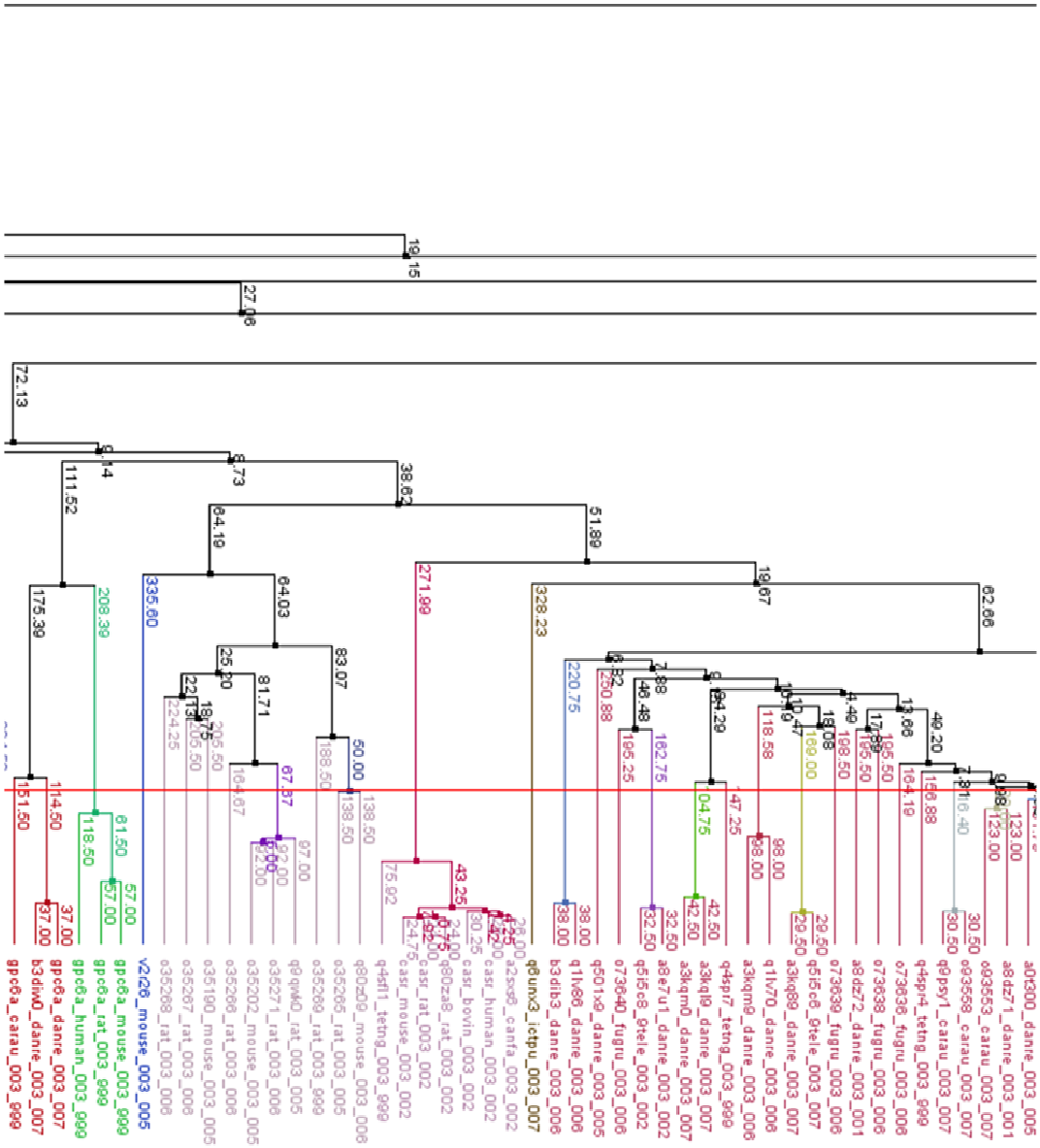
- continues on the next page -



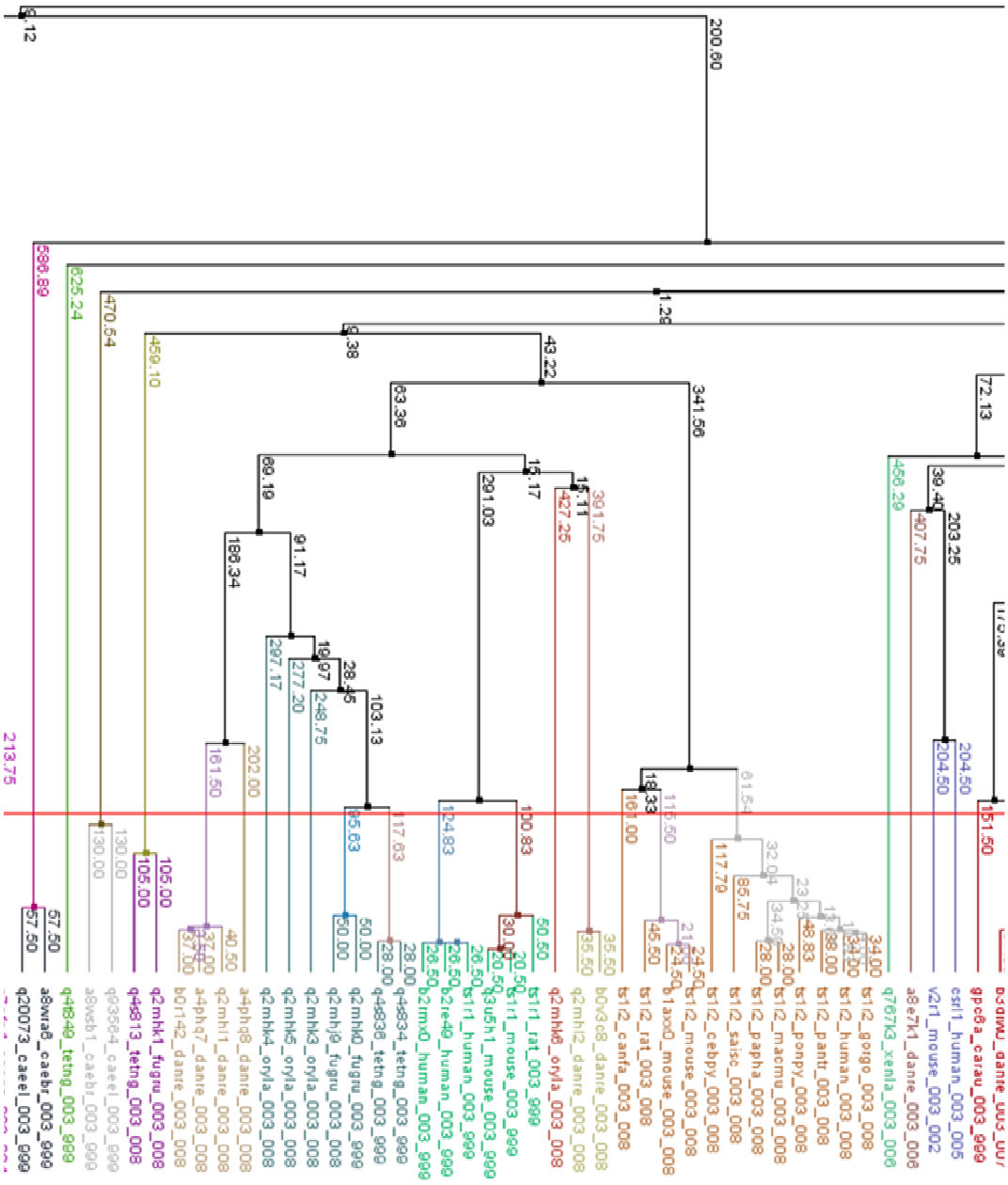
- continues on the next page -



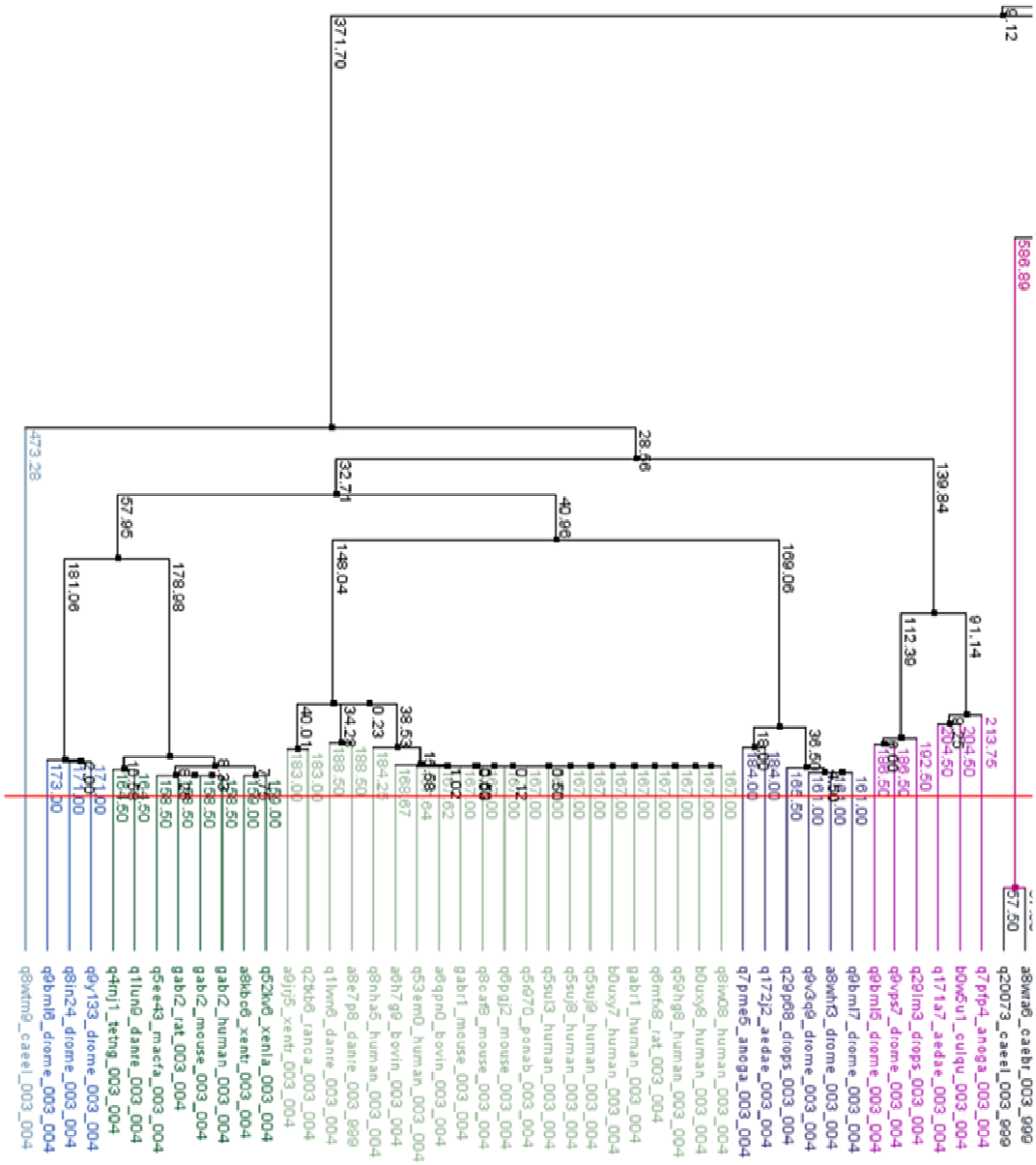
- continues on the next page -



- continues on the next page -



- continues on the next page -



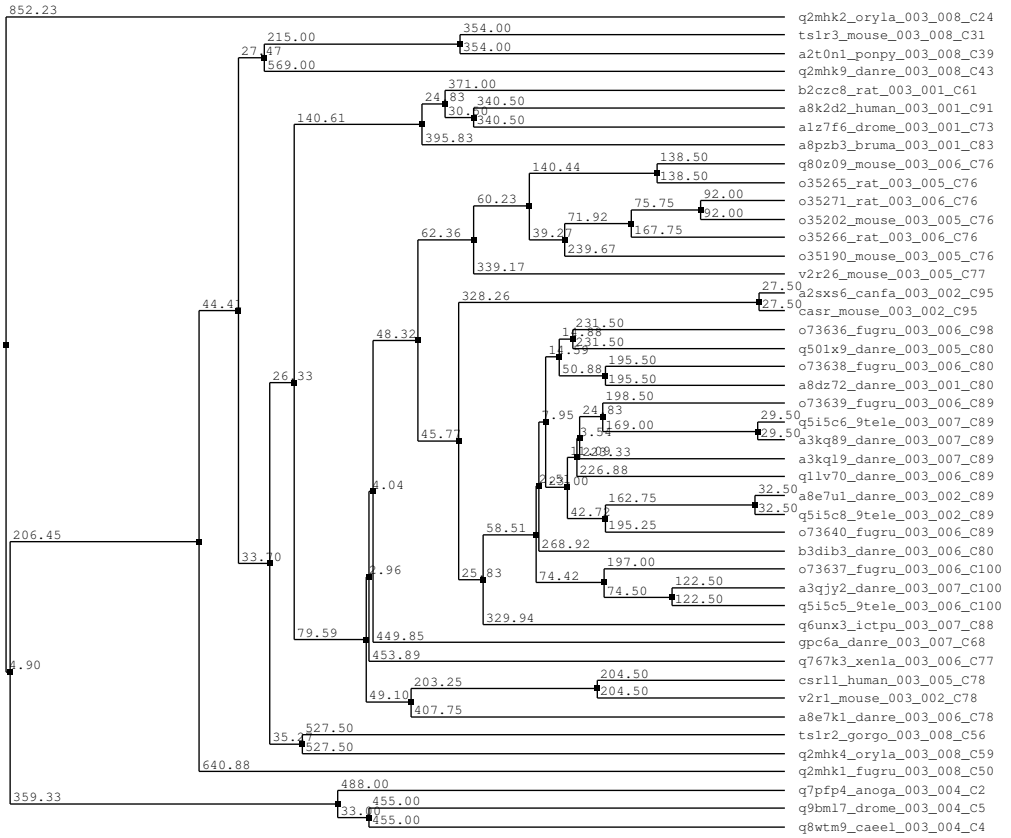


Figure B.10: The complete phylogenetic tree of figure B.9 is represented here in a reduced format in which the lower branches have been merged. The labels in the leaves are just representatives of the groups they summarize.